The Journal of

Economic Perspectives

A journal of the
American Economic Association

Winter 2014

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
No responsibility for the views expressed by the authors in this journal is assumed by the editors or by the American Economic Association.
THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Winter 2014, Vol. 28, 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 http://www.vanderbilt.edu/AEA/.
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 28 • Number 1 • Winter 2014

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 28, Number 1—Winter 2014—Pages 3–26

US Manufacturing: Understanding Its Past and Its Potential Future†

Martin Neil Baily and Barry P. Bosworth

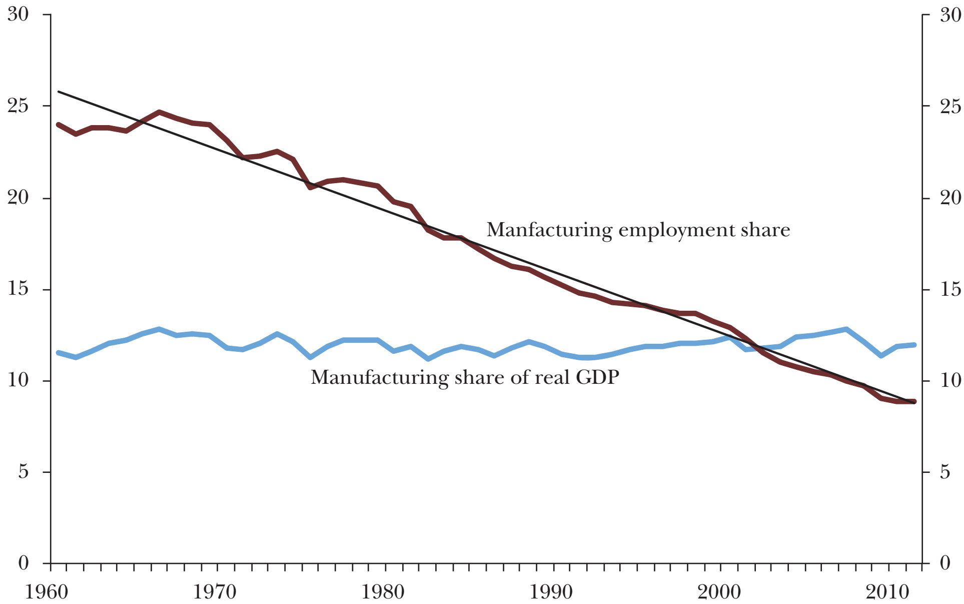
The development of the US manufacturing sector over the last half-century displays two striking and somewhat contradictory features: 1) the growth of real output in the US manufacturing sector, measured by real value added, has equaled or exceeded that of total GDP, keeping the manufacturing share of the economy constant in price-adjusted terms; and 2) there is a long-standing decline in the share of total employment attributable to manufacturing. These trends, going back several decades, are highlighted in Figure 1. Their persistence seems inconsistent with stories of a recent or sudden crisis in the US manufacturing sector. After all, as recently as 2010, the United States had the world's largest manufacturing sector measured by its valued-added and, while it has now been surpassed by China, the United States remains a very large manufacturer.

On the other hand, there are some potential causes for concern. First, though manufacturing's output share of GDP has remained stable over 50 years, and manufacturing retains a reputation as a sector of rapid productivity improvements, this is largely due to the spectacular performance of one subsector of manufacturing: computers and electronics. Meanwhile, the 90 percent of manufacturing that lies outside the computer and electronics industry has seen its share of real GDP fall substantially, while its productivity growth has been fairly slow. Complicating the

■ *Martin Neil Baily is a Senior Fellow in the Economic Studies Program and the Bernard L. Schwartz Chair in Economic Policy Development, while Barry Bosworth is a Senior Fellow in the Economic Studies Program and the Robert V. Roosa Chair in International Economics, the Brookings Institution, Washington, DC. Their email addresses are mbaily@brookings.edu and bbosworth@brookings.edu.*

† To access the disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.1.3 doi=10.1257/jep.28.1.3

Figure 1
Manufacturing Value Added and Employment as a Share of the Total US Economy, 1960–2011
(in 2005 prices)



Source: Industry Accounts of the Bureau of Economic Analysis.
Note: Output is measured as value added in 2005 prices, and employment is reported as persons engaged in production (full-time equivalent employees plus the self-employed).

matter, the data on output and purchased inputs suffers special measurement issues, raising questions about whether real output and productivity growth are overstated.

Second, although manufacturing's *share* of total US employment has declined steadily over the last 50 years (see Figure 1), recently there has been a large drop in the absolute *level* of manufacturing employment that many find alarming. After holding steady at about 17 million jobs through the 1990s, manufacturing payroll employment dropped by 5.7 million between 2000 and 2010. In large measure, the explanation lies with the equally striking decline of employment in the economy as a whole during the Great Recession and its aftermath, but the size of the absolute job loss deserves further examination.

Third, the US manufacturing sector runs an enormous trade deficit that had already reached \$316 billion by 2000, hit \$542 billion in 2005, and remains very high despite the recession, equaling \$460 billion in 2012; the manufacturing deficit is also very concentrated in trade with Asia, which represented over three-quarters of the deficit in 2000 and more than 100 percent in 2012. In 2000, only about one-third of the large deficit with Asia was accounted for by trade with China, but since then China has greatly increased its share, rising to 72 percent by 2012. The fall in manufacturing employment post-2000 has coincided with much of the growth in the bilateral trade imbalance with China, which suggests to some a causal link in which China trade is the reason for the loss of US manufacturing jobs. (And it may cause a feeling of *déjà vu* for those who remember the debate over trade with Japan in the 1980s and 1990s.) However, economists also recognize that trade imbalances are largely a macroeconomic phenomenon, reflecting the gap between national saving and domestic investment, so linking the trade imbalance to the problems of US manufacturing is more complex than blaming other countries.

In what follows, we examine each of these issues in greater depth and conclude with a discussion of the outlook for the future evolution of the manufacturing sector and its importance for the US economy. While US manufacturing remains an area of significant technological innovation, many of the largest US corporations continue to shift their production facilities overseas. It is important to understand why the United States is not perceived to be an attractive base for their production.

Trends in Manufacturing Output: Is It All about Computers?

Computers become faster and more powerful over time, implying large declines in their quality-adjusted prices.[1] Prices of computers, in this sense, do deflate in the ballpark of 10 or 20 percent a year. Thus we see much faster growth in the computer industry's real output than in its nominal output.

While the stability of the manufacturing share of *real* US GDP, highlighted in Figure 1, is striking, the manufacturing share of value added in *nominal* US GDP has fallen in half over the past half century—from about 25 percent in 1960 to 12 percent in 2010. The difference between these two perspectives is entirely due to the rapid fall in the quality-adjusted relative price of manufacturing output, which in turn is almost entirely driven by the fall in the quality-adjusted prices of computers and electronic products.

Pricing Computers

We have consistent industry-based data back to 1987, and over the 1987–2011 period, real value added in manufacturing expanded at almost the same rate as GDP as a whole, 2.6 as opposed to 2.5 percent per annum, as shown in Table 1. The exclusion of the computer and electronics industry, however, reduces the annual growth rate of US manufacturing output during these years to 0.6 percent. Even though the computer and electronics industry represented only about 10 percent of nominal value added in manufacturing, its real growth rate averaged nearly 20 percent annually—30 times that of the rest of the US manufacturing sector—and so its effects

[1] As an example, let's say that the price of one's new computer is the same as the price of one's old computer, but the new computer is four times better. Although the nominal price of the new computer is the same, in real terms, the new computer is only .25 times the price of the old computer.

Table 1
Annual Rates of Growth in Value Added of the US Manufacturing Sector and Growth in GDP, 1987–2011
(percentage change)

	1987–2011	*1987–2000*	*2000–2011*
Gross Domestic Product	**2.5**	**3.4**	**1.6**
Manufacturing	2.6	3.8	1.2
Manufacturing less Computers	0.6	1.5	−0.4
Durable Goods	4.0	5.4	2.5
Durable Goods less Computers	0.6	1.5	−0.5
Computers and Electronic Products	19.5	23.5	15.0
Nondurable Goods	0.7	1.4	−0.3

Source: Industry Accounts of the Bureau of Economic Analysis and authors' calculations.

on the growth of the overall manufacturing sector are dramatic. Table 1 also shows that the growth rate of production in durable goods as a whole averaged 4.0 percent annually from 1987 to 2011, comfortably exceeding the 2.5 percent annual rate of real GDP growth. But again if one looks at the growth of durable goods output while excluding computers and other electronics products, it is only 0.6 percent per year over this time. In short, outside of the computer and electronics industry, there is a clear decline in manufacturing's share of real GDP.

The measures of value added for the computer and electronics industry are heavily influenced by adjustments for the improved quality of these products. The measures are constructed by the Bureau of Economic Analysis within an input-output framework that provides nominal values for gross output of the industry, purchases of intermediate inputs, and value added. Indexes of the prices of gross output and intermediate inputs are assembled from the price index programs of the Bureau of Labor Statistics. The growth of real value added is the growth of gross output minus the growth of inputs, weighted by the nominal share of inputs in gross output.[2]

The relevant estimates of value added and its derivation in the computer and electronic products industry are summarized in Table 2. It is notable that the industry has expanded at relatively modest rates in nominal terms. Meanwhile, the rapid growth of real value added in the industry has been driven by the large declines in the quality-adjusted price index of its gross output. The nominal magnitude of purchased inputs has varied over time, but the prices of inputs have only slowly trended down: that is, the data imply that the rapid pace of innovation in computers and electronics is centered *within* the industry (not in the purchased inputs). Because value added has typically represented less than half of gross output, percentage changes in gross output imply even larger changes in value

[2] More specifically, the price of value added is an implicit price obtained from the ratio of the nominal and real values.

Table 2
Output Trends for the US Computer and Electronic Products Industry, 1987–2011

	Billions of dollars		
	1987	*2000*	*2011*
Value added	85.1	172.1	227.0
Gross output	216.4	503.6	350.1
Purchased inputs	131.4	331.5	123.1
Exports	18.8	55.5	48.4
Imports	14.8	89.8	119.7
Trade balance	4.0	–34.3	–71.3
	Annual percentage rates of change		
	1987–2011	*1987–2000*	*2000–2011*
Nominal values			
Value added	4.2	5.6	2.6
Gross output	2.0	6.7	–3.3
Intermediate inputs	–0.3	7.4	–8.6
Real values			
Value added	19.5	23.5	15.0
Gross output	9.0	15.4	1.8
Intermediate inputs	1.7	11.1	–8.5
Price indexes			
Value added	–12.8	–14.5	–10.8
Gross output	–6.4	–7.6	–5.0
Intermediate inputs	–1.9	–3.3	–0.2

Source: Industry Accounts of the Bureau of Economic Analysis.

added. Hence, a 6 percent average annual rate of decline in the price of gross output over the 1987–2011 period translates into a 13 percent rate of decline in the value-added deflator, and an extraordinary 20 percent annual rate of growth for real value added.

The domination of computers in the data raises some real concerns. First, it suggests that the seeming stability of manufacturing output (in value-added terms) may be more fragile than it appears, because it is being sustained by only a narrow part of the US manufacturing sector. Furthermore, this best-performing portion of manufacturing is rapidly moving overseas, and the United States now is a large net importer of computers and peripheral equipment, as shown in the top portion of Table 2. In two decades, the United States transitioned from being the global leader in producing computers to one among many players. That elevates the concern with sustainability. Second, while the technological advances within the computer industry contribute importantly to improvements in living standards through the expansion of computer services and the fall in their price, they may not raise the competitiveness of American workers because they can be incorporated with equal ease into foreign or domestic production of computers.

Input Price Bias

Another issue that arises with respect to the measurement of value added in the manufacturing industry is a concern that its growth may be overstated because of possible biases in the estimates of material inputs coming from mismeasurement of the input price deflators. The issue was explored in this journal by Houseman, Kurz, Lengermann, and Mandel (2011) in the context of offshoring of purchased inputs. If a US-based manufacturer shifts its purchases of components from one foreign producer to another in order to get a lower price, the components from the new source are treated in the price statistics as different products from those purchased from the previous supplier and the drop in their price is ignored. In the US consumer price statistics, the parallel issue is called "outlet substitution bias": that is, if consumers are buying the same product at a lower-priced warehouse store rather than at a higher-priced seller, it is not treated as a price decline, because prices are followed on a same-store basis. This problem arises in the price statistics even if the buyer switches suppliers within the United States, but it becomes more important as American companies start buying more low-price components from Asia or Mexico. It affects the measures of value added because if the increase in the price index for purchased inputs is overstated, the corresponding estimate of the growth in the real value of the inputs is understated, and the resulting measure of value added is too high.

Houseman et al. (2011) estimate that the growth in real value added of the manufacturing sector is overstated by about 0.2 percent a year. Feenstra, Mandel, Reinsdorf, and Slaughter (2013) draw on the economics of variety from the international trade literature and give a more expansive analysis of biases, one which takes outlet substitution bias into account but also estimates the impact of the increased number of suppliers now available.[3] They consider the whole business economy, not just manufacturing, and estimate that growth in GDP (real value added) originating in this sector is overstated by about 0.15 percent a year. However, most of this bias occurs in manufacturing. We think their findings imply about 0.7 percent a year overstatement of real value-added growth in manufacturing, coming roughly half from outlet substitution and half from the impact of the greater variety of input sources. However, as Feenstra et al. stress, it should be remembered that these estimates of bias in the measures of value added may to some extent be offset by other biases elsewhere.

The biggest potential bias in the input price data is in the computer and electronics industry, as both Houseman et al. (2011) and Feenstra et al. (2013) emphasize. Starting after 1997, the output price index for this industry was changed

[3] They also look at the impact of changes in tariffs and how they are treated in price and quantity indexes. They say these policy shifts did induce a bias in import price indexes, but only a very small one.

Table 3
Productivity Growth in US Computers and Manufacturing, 1987–2011
(average annual percent change)

	1987–2011	*1987–2000*	*2001–2011*
Output			
Total Nonfarm Business	2.8	3.7	1.7
Manufacturing	1.7	3.5	−0.4
Computers	8.0	14.6	0.7
Manufacturing less Computers	0.8	2.0	−0.5
Labor Productivity			
Total Nonfarm Business	2.2	2.0	2.4
Manufacturing	3.3	3.5	3.0
Computers	10.6	15.2	5.4
Manufacturing less Computers	2.3	1.9	2.8
Multifactor Productivity			
Total Nonfarm Business	0.9	0.8	1.1
Manufacturing	1.3	1.2	1.5
Computers	9.7	10.3	9.1
Manufacturing less Computers	0.3	−0.1	0.7

Source: Authors' calculations based on data of the Bureau of Labor Statistics Division of Major Sector Productivity, http://www.bls.gov/mfp/mprdload.htm.
Notes: Manufacturing industry data are based on gross output. Nonfarm business aggregates are based on value added.

from a "matched model approach," in which the prices of currently produced models are matched to similar models produced in the past, and shifted to an hedonic-based index, which seeks to estimate and quantify the quality change between models so that it can be counted as a rise in output. However, there was no corresponding shift in the methods used to construct the price index for inputs. The Bureau of Labor Statistics should take a hard look at its input price indexes and correct biases it can verify. However, multifactor productivity growth in the computer and electronics industry is estimated to be over 10 percent a year, fueled by the remarkable technological advances in this industry. Correcting the input price data and reducing estimated real value-added growth by roughly a percentage point will change the numbers, but it will not change substantially the basic story of the evolution of value added for the computer and electronics industry or for the rest of the manufacturing sector.

Productivity

The computer and electronics industry has a particularly large effect on evaluations of the productivity performance of the manufacturing sector. As shown in Table 3, labor productivity in total manufacturing advanced at an annual pace of 3.3 percent over the 1987–2011 period compared to 2.2 percent for the total nonfarm business economy. Similarly, multifactor productivity (MFP) appears to have grown more rapidly in manufacturing than in the overall business sector. Both of these outcomes seem very consistent with the popular view of manufacturing as a leading source of innovation and technological improvements. However, those beliefs appear to be driven in recent years by the performance of the computer and electronics industry, whereas the noncomputer manufacturing industries have rates of productivity improvement similar to or below the economy-wide average. As shown in the table, multifactor productivity improvements in the noncomputer manufacturing sector are particularly modest, averaging only 0.3 percent per year, compared to 0.9 percent for the overall business sector. However, capital–labor substitution and changes in the use of other inputs raise the rate of labor productivity increase to equal that of the nonfarm sector as a whole. In summary, the computer and electronics industry has a large impact on one's evaluation of the performance of manufacturing. This part of the sector has had tremendous quality-adjusted output and productivity growth, even allowing for data errors. In contrast, the noncomputer part of manufacturing has exhibited very slow output and multifactor productivity gains and only moderate labor productivity growth.

The Employment Decline in Manufacturing

The decline in manufacturing employment as a share of the economy-wide total is a long-standing feature of the US data and also a trend shared by all high-income economies. Indeed, data from the OECD indicate that the decline in the share of US employment accounted for by the manufacturing sector over the past 40 years—at about 14 percentage points—is equivalent to the average of the G-7 economies (that is, Canada, France, Germany, Italy, Japan, and the United Kingdom, along with the United States).

One explanation often given for this pattern is the "relative productivity hypothesis," which posits rapid relative productivity growth in manufacturing combined with unfavorable income and price elasticities. If output per worker rises more rapidly in one sector than in the rest of the economy, this will generally contribute to a fall in the relative price of that sector's output, which in turn will boost the demand for that sector's products. In addition, the overall increase in income in the economy as a whole coming from economy-wide productivity growth will add to sector demand. However, if the sector-specific price and income elasticities of demand are not large enough in absolute value, these sources of growth will not generate demand growth that exceeds the rate of productivity growth in that sector. The result is a trend decline in that sector's employment share. An appeal of the relative productivity hypothesis is that it could explain both the long-term trend decline in the manufacturing employment share in the United States and the fact that other advanced economies have seen the same pattern. Relatively rapid increases in output per hour and relatively low income and price elasticities could prevail both over many years and across countries. The evidence most often cited to support the relative productivity hypothesis is that labor productivity growth in manufacturing in the United States and in most other OECD economies has grown faster than overall labor productivity.[4]

The evidence on price elasticities is not as clear cut. When successful new products are introduced, the growth of demand can be very rapid if they achieve widespread acceptance. Once the market is "saturated," demand depends on the replacement cycle and growth slows. The introduction of the automobile before World War II is an example and so is electronics after the war. However, evidence from United States indicates that the manufactured goods share of the expenditure pie is falling over time measured in current dollars. Edwards and Lawrence (2013, table 3) document the fact that the share of consumption by Americans devoted to goods has declined from 50 percent in 1970 to 37 percent in 2000 and to 34 percent in 2010. Current-dollar US fixed nonresidential investment has also declined as a share of GDP since the 1980s (Baily and Bosworth 2013). (Note that the expenditure shares include manufactured goods purchased from overseas.)

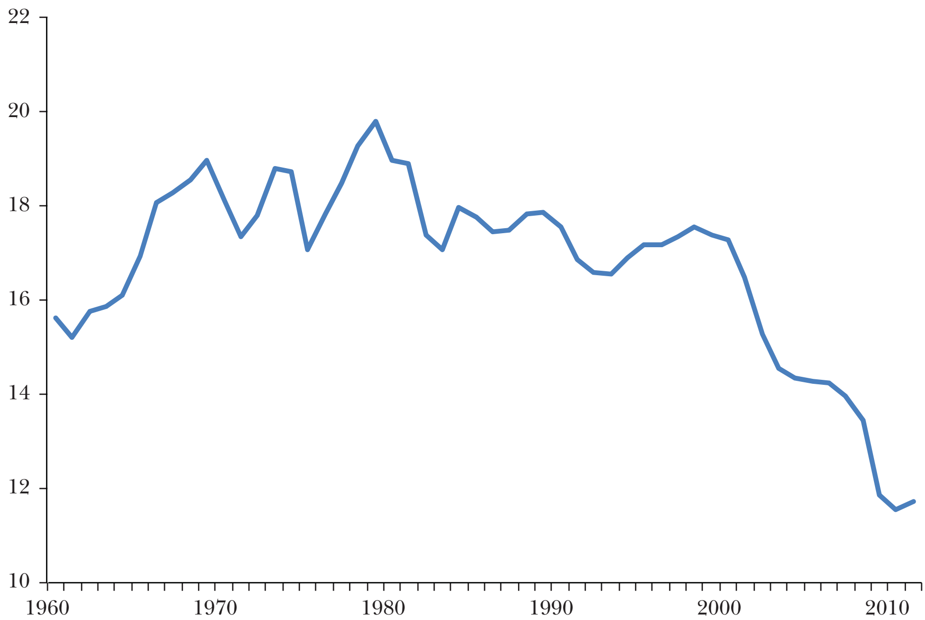
The relative productivity hypothesis runs into trouble, however, in explaining what has happened in the last 20 years or so. Productivity in the noncomputer segment of US manufacturing rose by 2.3 percent a year from 1987 to 2011 compared to 2.2 percent a year in the nonfarm business sector as a whole, as shown in Table 3. The labor productivity growth rates are the same, within a tenth of a percent. Whatever may have been true in the past or in other countries, the falling share of US employment in manufacturing excluding computers was not driven by excess productivity growth in 1987–2011. (In contrast, rapid relative productivity growth was important in the computer and electronics industry, where productivity growth was extraordinary.) The continued trend decline in the manufacturing employment share since 1987, therefore, was the result of the falling share of spending going to manufactured goods and the falling fraction of those goods that were produced in the United States.

The dramatic fall in the absolute level of US employment in manufacturing after 2000—a loss of one-third of the total by 2011—has been noted by many observers. It seems to conflict with Figure 1, which shows that manufacturing's employment *share* has been declining and continues to decline at a steady rate. By itself, the change in manufacturing employment, shown in Figure 2, certainly seems unprecedented. After remaining flat throughout the 1980s and 1990s, the total number of manufacturing jobs fell continuously in the 2000s. The largest decline is in computers and electronics, but the job losses are large in all of the underlying industries.

However, what is missed in a focus on a single sector is that job weakness after 2000 was not just a manufacturing issue; employment in the entire US economy went through a negative shift after 2000. Comparing the period 2000–2011 with

[4] The major source of international data on productivity is the Structural Analysis (STAN) Statistics of the OECD at http://www.oecd-ilibrary.org/industry-and-services/data/stan-oecd-structural-analysis-statistics_stan-data-en. Prior to its elimination in the 2013 budget cuts, the International Labor Comparisons program of the US Bureau of Labor Statistics also provided comparable international data. For instance, http://www.bls.gov/news.release/prod4.nr0.htm.

Figure 2
Persons Engaged in Production in US Manufacturing, 1960–2011
(millions)



Source: Industry Accounts of the Bureau of Economic Analysis.
Note: Persons engaged in production are measured as full-time equivalent employees plus the self-employed.

the period 1987–2000, we see that the average annual growth rate of total nonfarm employment slowed from 2.1 percent per year to –0.1—a deceleration of 2.2 percent per year—compared to a 3.3 percentage point deceleration in manufacturing.

After a sluggish recovery in the first couple of years after the 1990 recession, US aggregate demand started to surge as consumption grew rapidly, the technology boom spurred investment, and residential and commercial construction increased. Employment in all sectors was the beneficiary of this domestic demand growth, as actual GDP went well above potential GDP, according to the Congressional Budget Office. US manufacturing employment was sustained by the boom, remaining flat during the decade through 2000 despite rapidly increasing imports and a rising trade deficit (which reached $316 billion in 2000). After 2000, employment in the whole economy grew much more slowly and was ravaged in the Great Recession. Most major sectors of the economy experienced negative job growth or a substantial slowing, but manufacturing was among the hardest hit. Since 2000, there has been no significant change in relative rates of productivity growth between manufacturing and the total economy after 2000, so the relative decline of manufacturing employment is also reflected in output growth. Edwards and Lawrence (2013) attribute that outcome to a high elasticity of manufacturing output to fluctuations in aggregate demand. Others, however, point to increased import penetration as a major contributor (Pierce and Schott 2012).

The Role of Trade and the Importance of China

The US economy last achieved a balance of trade in manufactures in the early 1980s, and the size of the manufacturing trade deficit has grown steadily over time. In the mid-2000s, the size of the US trade deficit in manufacturing was equal in value to nearly half of manufacturing value added and is now equivalent to about 40 percent. Furthermore, the US deficit in manufacturing trade of $460 billion in 2012 exceeded the total current account imbalance of $440 billion. (The US economy has a surplus in services trade and records a positive and rising net inflow of income on its foreign investments, despite being a large net debtor.) These statistics are used to argue that the rising trade deficit—and in particular the trade deficit with China—is the primary cause of the sharp decline in manufacturing value added and employment after 2000. Pierce and Schott (2012), for example, link manufacturing employment loss to the granting of permanent normal trade relations to China in 2000. They report that "industries where the threat of tariff hikes declines the most experience greater employment loss due to suppressed job creation, exaggerated job destruction and a substitution away from low-skill workers" (quoted from the abstract). Autor, Dorn, and Hanson (2013) estimate that China's improved competitive position between 1991 and 2007 explained at least 25 percent of the decline of US manufacturing employment during that period, about 40 percent during the 2000s. Not everyone agrees that trade is important for the post-2000 employment decline. For example, Edwards and Lawrence (2013) point out that the trade deficit was already very large in 2000, and they estimate that the job content of the trade deficit in 2012 was very similar to what it was in 2000. They conclude that trade accounts for little of the fall in employment after 2000.

It is notable that for most of the past three decades, a growing trade deficit was associated with a buoyant domestic economy, rapid job growth, and a decline in unemployment to unprecedented levels.[5] This domestic strength suggests that the trade deficit was not something forced on the US economy by outside pressures, but rather a response to changing domestic economic conditions that pushed aggregate demand beyond the nation's productive capacity. The excess demand was satisfied in a noninflationary way by exporting less and importing more. This was reflected in turn by an increase in foreign financial investments in the United States coming from a rise of saving relative to investment in other countries, a large growth of supply capacity in countries that export to the United States, and a stable or rising value of the dollar. Many of the discussions of the role of trade on the evolution of the manufacturing sector ignore that macroeconomic context.

[5] Using the Congressional Budget Office's measure of potential GDP, the utilization rate for total GDP averaged 100 percent over the 1985–2007 period.

Table 4
US Balance in Manufactures Trade by Area, 2000–2012
(billions of dollars)

Item	*2000*	*2005*	*2012*	*Change 2005–2012*
Total	**–316**	**–542**	**–460**	**83**
Asia	–240	–372	–478	–106
China	–84	–206	–342	–137
Hong Kong	3	7	31	23
Other Asia	–160	–173	–166	7
Canada	–15	–16	46	62
Latin America	–3	–28	57	85
Europe	–58	–131	–128	3
Middle East & Africa	1	4	43	39

Source: US Department of Commerce, Bureau of the Census, "International Trade in Goods and Services, Supplementary Tables," various years.

Between the early 1980s and the end of the boom in 2007, Americans devoted ever-increasing shares of their incomes to consumption—the consumption share of GDP rose by 6 percentage points to 67 percent by the end. The surge of consumer spending was reflected in a substantial reduction in private saving. At the same time, the boom in information technology made the United States a particularly attractive location for business investment, and a strong expansion of residential investment contributed to the growing domestic imbalance between saving and investment.

These trends have large implications for manufacturing because it dominates the tradables sector of the economy: thus, the emergence of sustained trade imbalances will lead to major shifts in the size and composition of the domestic manufacturing sector. In the presence of sustained domestic demand, firms are often content to focus on domestic markets, and those that wish to increase their global sales will do so through the expansion of their overseas production facilities rather than exporting out of a fully employed domestic economy.

Of course, all of this has changed in the aftermath of the financial crisis. The United States is faced with considerable unemployment and may be stuck on a low-growth path for some time to come. With deficient domestic demand, it is now very interested in improving its trade performance as a means of reviving the domestic economy, but it is hard to reverse course in a weak global economy.

As shown in Table 4, the United States has typically had a large manufacturing trade imbalance with most regions of the world. The total has fallen from its peak in the mid-2000s, but the imbalance with Asia continues to increase. The deficit with China is particularly striking both in its size and its rate of increase in recent years. In general, economists prefer a multinational perspective on trade as opposed to an emphasis on the bilateral relationships, but over the past decade the magnitude of the US–China bilateral imbalance has reached extreme levels. China's trade surpluses reflect its own macroeconomic imbalances: an excess of national saving over domestic investment (even though both seem extraordinarily high by American standards).

It is instructive to divide China's trade regime into two distinct components, normal trade and processing trade, because the two have been evolving in different ways. About half of China's trade is accounted for by processing activities, which are based on the duty-free import of goods to be assembled and re-exported. The distinguishing features of processing activities are the low contribution of domestic value added and domination by foreign-invested enterprises (accounting for about 80 percent of output). In fact, China's processing trade is an integral part of a larger regional production network of companies in Asia, companies that had long exported to the United States and have now moved their assembly work to China. Morrison (2012) shows that US imports from the Pacific Rim countries, including China, have been a nearly constant share of US manufacturing imports since 1990, but that China's share of that trade rose from 8 percent in 1990 to 55 percent by 2011. The exports from these countries are widely diversified by recipient country, but the United States is the largest single destination.

The development of transnational production networks is but one of a series of profound organizational innovations that have given millions of poor unskilled workers—particularly in Asia—access to the global economy. Capital and technology are now mobile in ways they never were before, and both can move about the globe in search of the optimal combinations of skilled and unskilled labor and preferred institutional arrangements. The result has been an unprecedented growth of a global middle class. But those innovations have also introduced a rapidly changing set of circumstances for workers and firms in developed economies that are now exposed to much more intense global competition.

The distinction between the processing and normal trade components is important for evaluating the importance of trade as a driver in China's overall growth. China's trade sector is certainly oversized by the standards of other large countries; but, as highlighted in a recent paper by Koopman, Wang, and Wei (2012), the domestic value-added content of processing exports from China is much lower than that of normal exports. They used a detailed input-output table to estimate the foreign and domestic content of China's exports, and found a sharp contrast in the 1990s when the domestic content of China's processing exports was only about 20 percent compared to 90 percent for normal exports. The processing and normal trade exports have become more similar over time: the domestic share of China's processing goods has steadily grown as the foreign firms have increased their reliance on local sources for the components, and the domestic producers of normal exports have increased their use of foreign inputs. However, the domestic content of processing exports is still less than half that of normal exports.

China's normal (nonprocessing) trade has also grown very rapidly, and processing trade is actually a shrinking share of the total, falling from about 55 percent of the total in 2004 to about 44 percent in 2011 (based on data from China Customs). China's balance of normal trade has fluctuated over the years, and it has been in substantial deficit since 2008, as shown in Figure 3. As a result, processing trade currently accounts for China's entire trade surplus.

China's processing trade and the growth of the Asian production network are of particular importance to an understanding of the evolution of the US manufacturing sector. Many American firms have shifted away from the prior model of large integrated production units to focus on product design and marketing. They contract with firms that are part of the regional production network in Asia, and undertake little of their own production. For example, the US trade deficit in the computer and electronics industry rose from 14 percent of gross industry output in 1998 to 56 percent in 2011, and the industry accounted for 38 percent of all imports from China in 2012. Apple Inc. is a leading example of such a company: in recent years it has owned no large production facilities in the United States or elsewhere, preferring to contract with companies based in Taiwan and Korea who assemble the products in China (like Foxconn, for example). But by controlling key elements in the value chain, Apple extracts much of the profit. Similar networks have become common in the market for personal computers. In contrast, Mattel has also closed all of its production facilities in the United States, but continues to operate factories throughout Asia.

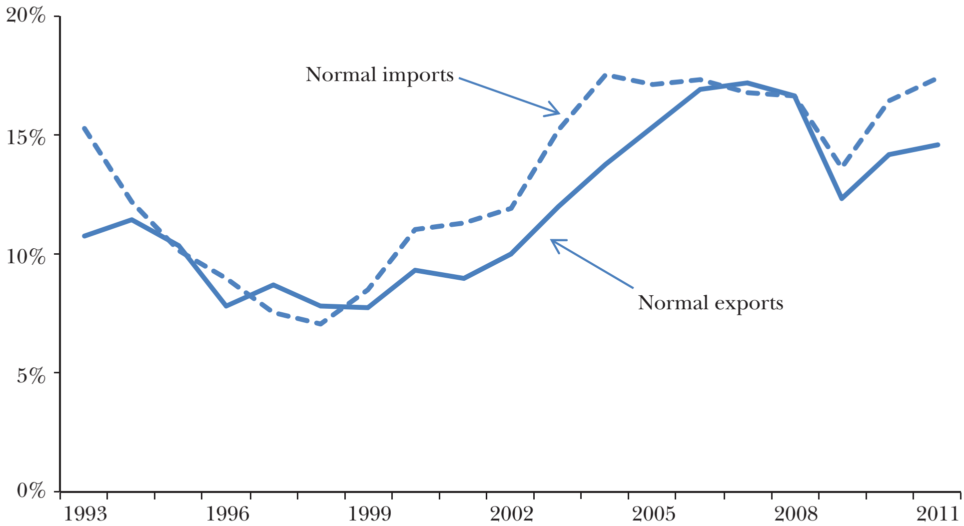
Most of the analysis of the US trade deficit with China has focused on the import side of the accounts. Yet measured as a share of the importing country's GDP, the magnitude of US manufactured imports from China is virtually the same as the magnitude of Europe or Japan's imports from China.[6] In 2010, imports from China accounted for 2.6 percent of GDP for the EU-15, 2.8 percent in Japan, and 2.7 percent for the United States. Instead, the differences in trade are on the export side, where exports to China account for only 0.8 percent of US GDP, compared to 1.2 percent for Europe and 3.5 percent for Japan. Given the historical antagonism between Japan and China, it is difficult to accept the view that China discriminates against the United States to a greater degree than Japan. Of course, Japan is closer to China; but, in a broader context, where we have used gravity equations to compare US, EU, and Japanese trade flows with a large number of trading partners, the US share of imports in GDP is normal, but its export share is consistently far below that of both Europe and Japan (Bosworth and Collins 2010).[7]

The United States is an important but not dominant export market for China. Exports to the United States were 20 percent of China's total in 2011, about the same as the US share of global GDP, but this was down from 29 percent in 2001. Furthermore, while China has a large trade surplus with the United States, many of its exports are in the processing sector where the value-added benefits to China are limited. On the export side, China is America's most rapidly growing market,

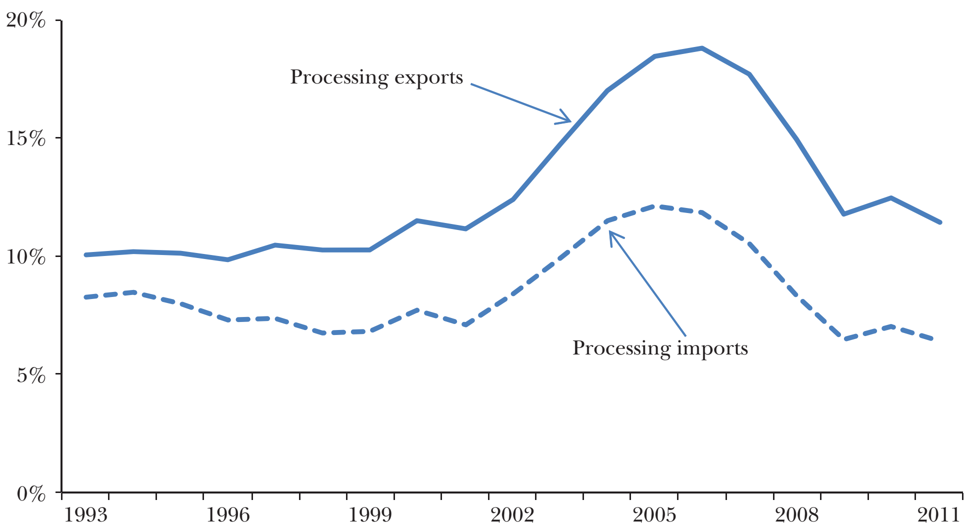
[6] To obtain a consistent classification, over time, we defined Europe as the aggregate of the original 15 members of the EU and excluded intragroup trade.
[7] Gravity equations are a common means of benchmarking trade flows. The volume of bilateral trade is related to the GDP/capita, population, and distance between trading partners.

Figure 3
Components of China's Merchandise Trade, 1993–2011
(percent of GDP)

A: Normal Exports and Imports



B: Processing Exports and Imports



Source: China Customs.

but it is important that the US share of the Chinese import market declined from 11.4 percent in 2001 to 7.7 percent in 2011 (calculated from IMF, *Direction of Trade Statistics*).

Even though Edwards and Lawrence (2013) argue that trade does not explain the post-2000 drop in manufacturing employment, they do show that the trade deficit does have a large jobs component, and its elimination would raise manufacturing employment by about 25 percent. During the boom years, Americans were relatively indifferent to the consequences of large trade deficits, but in the future, the United States cannot afford ongoing trade deficits of 3 percent and more of its GDP. Because of the domination of manufacturing in tradables, much of any adjustment will have to be concentrated in that sector. We now turn to a discussion of the future of US manufacturing and policies that would assist in reducing manufacturing trade deficits.

The Future of Manufacturing

Historically, the share of manufacturing employment in the total has declined by about 0.3 percentage points a year. The Congressional Budget Office projects civilian employment to be 159 million in 2023, so if the historical trend were to continue into the future, the implied level of manufacturing employment would be about 10 million, or about 2 million jobs below its current level. Although keeping this historical trend in mind offers a useful perspective, it is of course not set in stone, and there are more optimistic projections. One view is that market forces and new technologies are already aligned to bring about a re-shoring revolution, bringing jobs and production back to the United States. A somewhat different view is that the right set of government policies could generate a different future path.

In this section we look first at the importance of manufacturing to the economy. Is it special and does that mean special policies are needed? We look at how new technologies may impact manufacturing, and then at what policy changes might both help manufacturing and be justifiable economically.

Why is Manufacturing Important?

Any large sector of the economy deserves study, but manufacturing also has characteristics that make it of special interest. First, Americans live and work in a global economy in which they exchange products that they produce for those that they consume (and they wish to do so without a continuous decline in their terms of trade). Manufactures account for a very large proportion of tradables, and a large proportion of the industries in which the United States could potentially run a trade surplus. While the US economy does have a trade surplus in many services industries, it is not nearly substantial enough to offset continued large deficits in goods trade. Thus, in the context of the global economy, the United States must become a better exporter, and realistically that means more exports of manufactured products.

Second, a strong domestic manufacturing sector offers a degree of protection from international economic and political disruptions. This is most obvious in the provision of national security, where the risk of a weak manufacturing capability is clear. Overreliance on imports and substantial manufacturing trade deficits increase Americans' vulnerability to everything from exchange rate fluctuations to trade embargoes to supply disruptions from natural disasters.

Third, manufacturing is often identified as an area in which much of the country's research and development takes place, but as illustrated by industries such as pharmaceuticals and consumer electronics, manufacturing has become increasingly separate from research and development. Also the manufacturing that remains has become increasingly capital-intensive and the US economy performs well in the production of machinery for manufacturing.

Fourth, the long-term decline in the manufacturing share of employment has meant fewer jobs available at good wages for workers who lack advanced education (for example Goldin and Katz 2008). The loss of nearly 6 million jobs since 2000 has been damaging to workers who have been laid off, communities that have lost a vital source of employment, and to young workers who might have found jobs in the sector (Autor, Dorn, and Hanson 2013).

We agree with those who think manufacturing is important, but we do not agree that this justifies special treatment for the sector, such as special tax rates or other subsidies. Such policies are hard to enforce and an invitation to arbitrageurs who seek ways to capture the gains from such subsidies with little or no actual change in real behavior. However, certain policy decisions have had a particularly large adverse impact on manufacturing because it is so exposed to global competition. We will discuss these below.

Emerging Technologies in Manufacturing

The future of American manufacturing will largely be determined by the extent to which it can take advantage of various new technologies that will influence the structure of manufacturing in future years (for example, Hart, Ezell, and Atkinson 2012; McKinsey Global Institute 2013). There is at least suggestive evidence that technological innovation is continuing at a rapid rate; for example, the rate of issuance of patents to US residents has increased substantially since the 1970s, based on data from the US Patent and Technology Office (although there are problems of comparability over time with changing standards of what can be patented and how quickly patents are processed). In this section, we describe some of the most important emerging technologies in manufacturing: industrial robots and automation; additive manufacturing; advanced design; direct interconnections over the Internet between sensor and machines; materials science and biotechnology; and energy production. While these technologies have the potential to increase US-based output in many sectors of manufacturing, the number of net new jobs that would be created remains an open question.

The last few years have seen rapid strides in the technology of *industrial robots and automation*, allowing robots to perform tasks that can today only be performed by humans. Many of these tasks require dexterity that robots are only now acquiring, while others require minor adjustments and variances, which are difficult to program a machine to respond to. While industrial robots have been used in several industries for heavy lifting, dangerous operations, and repetitive, precise movements—painting and welding in the auto industry, for example—they have been priced well out of range for more regular "human" tasks. That could change soon with the development of robots that have the capability to work safely alongside humans. For example, a robot priced at $20,000 can now sense a human in the path of its arms and stop movement. It can be "reprogrammed" for new tasks by a human operator who physically manipulates its arms to move, bend, lift, or drop in the desired way. These low-cost robots have the potential to increase precision and raise productivity by reducing the number of workers required. Of course, robotics is a two-edged sword for US manufacturing employment. These advances could reduce the number of jobs for a given level of output, but at the same time, the advantage in labor costs currently held by Asian assembly and manufacturing companies would be reduced or eliminated, allowing production to be re-shored to the United States.

Additive manufacturing refers to a range of technologies, including 3D printing, that build up objects from small particles. Thus far, 3D printing has been used primarily to create prototypes or objects that would be impossible to machine; but in the future, companies will sell designs on the web, instead of selling products directly. Customers will be able to print out the desired product for themselves, or companies will provide 3D printing services on a contract basis. As the technology improves further, some products will be customized to match the specific demands of the individual customers. Additive manufacturing can increase flexibility, cut development costs and time, reduce material waste, eliminate tooling costs, and simplify production runs. How fast this technology will be deployed is hard to predict. But even now, when additive manufacturing is still expensive, it is becoming a standard tool for rapid prototyping and early production runs for small and complex components.

Increases in computer power and advances in software are leading to *advanced design*: that is, enhancing companies' ability to develop digital prototypes and carry out much more testing on the digital model before building a physical prototype. McKinsey Global Institute (2013) estimates there will be a 20–50 percent reduction in research and development costs as well as reduction in time to market.

The "Internet of Things" refers to how low-cost sensors will lead to a widespread array of *direct interconnections over the Internet among machines and locations.* Connecting machines will allow improved monitoring of production processes remotely and allow operators to provide instructions to one set of equipment based on activity at other equipment. Process designers will be able to set up systems that automatically make adjustments based on sensor readings from all the equipment in a network—and then apply optimization algorithms to improve efficiency. One specific example is to reduce energy usage. Traditionally, motors operate at peak capacity irrespective of load. However, smart motors are able to adjust power usage as output changes, usually through variable speed drives controlled by an intelligent motor controller. With low-cost sensors allowing improved inter-machine and system communication over wireless networks, it will be possible to make manufacturing systems that include thousands of smart motors, enabling substantial improvements in energy efficiencies in manufacturing. Connecting machines within a given factory, or even across multiple factories, will allow particular machines or conveyor belts to be shut down when not in use, saving energy and wear.

There have been breakthroughs in *materials science and biotechnology* that promise major advances ahead, although the timetable for adoption is unclear. Applying the technology to carbon nanotubes and graphene has allowed the creation of high-performance transistors and ultra-strong and light composite materials. Fluorescent nano-particles are used in biological labeling and solar cells. In biotechnology, nanoenabled technologies allow more rapid diagnosis of illnesses, detect contaminants, and provide glucose monitoring and many other applications. Bringing these advances into the economic mainstream will require long time horizons and continued investment.

Finally, although the recent developments in energy production technology are not manufacturing innovations, the extraction of natural gas and light tight oil from shale deposits will have a substantial effect on manufacturing. US natural gas resources have nearly doubled since 2003, driven by the development of shale deposits nationwide. The United States has the second-largest recoverable shale gas reserves in the world at 24 trillion cubic meters (after China's estimated reserves of 36 trillion cubic meters). However, the United States is substantially ahead of the rest of the world in having started to tap these reserves at increasing scale. By 2020, shale gas is expected to grow to over 25 percent of total natural gas production, which will lead to a 60 percent drop in natural gas imports. On net, US energy imports could fall to zero. Production of "light tight oil" (LTO), which also comes from fracking, has also developed rapidly. Current LTO production estimates for 2020 are between 5 and 10 million incremental barrels per day. US consumption of oil reached around 20 million barrels per day before the recession and has been around 19 million barrels per day since then (based on data from the Energy Information Administration). This new wave of energy production does raise legitimate environmental concerns, but it should be possible to develop the oil and gas fields responsibly.

Because natural gas is expensive to ship around the world and exports are restricted by regulation, the expansion of the North American supply will drive US prices for natural gas below world levels: for example, natural gas may very possibly be priced at $4–6 per million BTUs in the United States, well below the $12 price range in Europe and $16 in Asia. (Oil prices are set globally, but it is likely that US domestic prices will carry some differential below imported oil and that the greater security of domestic supply will be an attraction for users. West Texas Intermediate crude has been priced $10 to $15 a barrel below the Brent crude benchmark of international crude oil prices for much of 2013.) Cheaper natural gas will also keep electricity prices down. The cost of new capacity using natural gas turbines is estimated to be about 4 cents per kilowatt/hour at today's prices, compared to 6 cents for new coal-fired capacity and over 10 cents for nuclear or solar power.

Manufacturing tends to be an energy-intensive sector, and thus it benefits from the prospect of less-expensive energy supplies. Indeed, both US-based and global companies are already investing in new plants in the United States to take advantage of the lower price of energy and natural gas as a feedstock. PricewaterhouseCoopers (2012) suggested that the movement of energy-intensive sectors such as chemicals and plastics back to the United States would result in about one million new jobs in manufacturing, although this figure may be overoptimistic. After all, the most energy-intensive industries tend also to be the most capital intensive and have relatively low levels of employment. Additional manufacturing employment will come from the development of the new energy sources. McKinsey Global Institute (2013) estimates that exploiting the domestic oil and gas will require capital investment of $1 trillion over the next five to ten years. Another upside for the manufacturing sector from low energy prices is the potential for the US economy to shift part of its transportation system to natural gas, which would generate substantial demand for manufactured products. Some companies such as GM, Navistar, and Cummins are developing natural gas–powered trucks (Smith 2012), and in turn, some companies are already shifting their short-haul trucks to natural gas because of low fuel cost, a trend that could extend to long-haul trucks also if the refueling infrastructure is developed.

The main potential downside to US-based manufacturing from the increase in domestic energy production lies in a version of the "Dutch disease," which in this case refers to the way in which reducing net energy imports might raise the exchange rate and make domestic manufacturing less competitive.

What Role for Public Policy and Manufacturing?

Given the importance of the manufacturing sector, what policies will give growth and employment in this sector the best chance in the future? While we do not support special subsidies, it is important to ensure that existing policies are supportive of manufacturing—or at least do not discriminate against it. Our policy recommendations are all intended to make the United States a more attractive location for manufacturing production.

The macroeconomic factors creating the US trade deficit put US manufacturing at a disadvantage by distorting capital flows and the foreign exchange value of the dollar. The US economy has experienced an unusual confluence of factors that encouraged Americans to consume beyond their means financed by the steady sale of assets to foreigners in return for a large net inflow of imports. Going forward, consumption expenditures and investments in residential housing will constitute a smaller share of GDP. As of 2013, the US economy is making little progress in filling an ongoing production gap of about 5 percent of potential GDP. Yet consumption remains at an elevated share of GDP, financed by large but ultimately unsustainable fiscal transfers. Historically, the external trade deficits have been sustained by a shortfall of domestic saving compared to domestic investment, and it seems clear that insufficient levels of national saving drove up the exchange rate and priced US exporters out of foreign markets. The government has few if any tools by which it can dramatically raise private saving; thus, the increment to national saving will be achieved most effectively by steps to reduce the federal budget deficit.

This policy recommendation, which we have made forcefully in the past, is problematic at present because of the weakness of aggregate demand both in the US economy and around the world. The first priority of macroeconomic policy is to help restore full employment, and that goal would also help manufacturing. Fiscal consolidation in the very near-term could abort the recovery. Over the longer term, however, it is hard to see how the United States can significantly improve its trade balance without tackling the budget deficit.

We noted earlier that the US trade deficit is driven by low exports as much or more than high imports. We believe that US exports could benefit from trade negotiations that help to pry open foreign markets, and by negotiations with countries that manage their exchange rates about the appropriate level of their exchange rates.

Another important step that the US could undertake to become a more attractive location for manufacturing in a world of global supply chains involves its tax code. The marginal rate of corporate taxation in the United States is too high, particularly in relationship to the tax rates of other countries. In a world economy where choices about capital, technology, and production facilities are increasingly flexible, this is inducing firms to locate overseas. The United States has the highest corporate tax rate within the OECD, and, at a combined 39 percent, it exceeds the average by 14 percentage points. The United States needs to follow the lead of other countries in shifting toward greater reliance on consumption-based taxation.[8]

Both American companies and foreign companies investing in the United States report that the skills of the US workforce are comparatively weak. It lags behind many other countries in developing effective vocational education and job-training programs, and the educational attainment of young workers is falling behind that of countries like Canada, Japan, and Korea. Furthermore, US 15-year-olds rank 25th in math and 17th in science in PISA (Program for International Student Assessment) scores among OECD nations. Germany is an example of a country that has used a high-quality vocation education system to improve the skills of its workforce. Greater attention needs to be paid to reversing the deterioration in US workforce skills.

Similarly, the United States suffers from a deteriorating physical infrastructure that raises the costs of production and limits the location of export activities. The extraordinarily low level of current interest rates suggests that now is an ideal time to engage in long-term borrowing of funds to finance the repair and modernization of those systems. Such financing, if matched by a credible dedicated revenue source

[8] The United States also attempts to tax the foreign income of US companies, albeit with a deferral. Most other countries use a territorial-based system in which income is taxed only in the country in which it is earned.

for repaying this debt, would not add to concerns about an unmanageable level of general fund debt.

Supporters of manufacturing often stress the need for government support of technology (Atkinson, Stewart, Andes, and Ezell 2012). We agree, but as described earlier, US companies remain strong in technology development while viewing it more profitable to produce the technology goods overseas.

Manufacturing is an important sector and needs to be competitive in order for the US economy to return to full employment with a sustainable trade balance. But improving the US export position cannot be quickly achieved. It will take years to rebuild the domestic supply chain and undo the incentives that have encouraged American firms to shift their production abroad. The key to expanding US exports and reaching manufacturing's employment potential is to have companies, domestic and foreign, judge it is profitable to manufacture here.

■ *The authors would like to thank the editors of this journal for many helpful comments.*

References

Atkinson, Robert D., Luke A. Stewart, Scott M. Andes, and Stephen J. Ezell. 2012. *Worse than the Great Depression: What Experts Are Missing about American Manufacturing Decline.* Washington, DC: Information Technology and Innovation Foundation.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103(6): 2121–68.

Baily, Martin Neil, and Barry Bosworth. 2013. "The United States Economy: Why Such a Weak Recovery?" Paper prepared for the Nomura Foundation's Macro Economy Research Conference, "Prospects for Growth in the World's Four Major Economies," September 11, 2013. http://www.brookings.edu/research/papers/2013/09/united-states-economy-why-weak-recovery-baily-bosworth.

Bosworth, Barry, and Susan Collins. 2010."Rebalancing the U.S. Economy in a Post-Crisis World." Presented at the ADB-Brookings Conference on "Trans-Pacific Rebalancing," held March 3–4, 2010, in Toyko.

Edwards, Lawrence, and Robert Z. Lawrence. 2013. *Rising Tide: Is Growth in Emerging Economies Good for the United States?* Washington, DC: Peterson Institute for International Economics.

Feenstra, Robert C., Benjamin R. Mandel, Marshall B. Reinsdorf, and Matthew J. Slaughter. 2013. "Effects of Terms of Trade Gains and Tariff Changes on the Measurement of US Productivity Growth." *American Economic Journal: Economic Policy* 5(1): 59–93.

Goldin, Claudia, and Lawrence F. Katz. 2008. *The Race Between Education and Technology,* Belknap Press.

Hart, David M., Stephen J. Ezell, and Robert D. Atkinson. 2012. "Why America Needs a National Network for Manufacturing Innovation." Information Technology and Innovation Foundation. http://www2.itif.org/2012-national-network-manufacturing-innovation.pdf.

Houseman, Susan, Christopher Kurz, Paul Lengermann, and Benjamin Mandel. 2011. "Off Shoring Bias in U.S. Manufacturing." *Journal of Economic Perspectives* 25(2): 61–80.

Koopman, Robert, Zhi Wang, and Shang-Jin Wei. 2012. "Estimating Domestic Content In

Exports when Processing Trade Is Pervasive." *Journal of Development Economics* 99(1): 178–189.

McKinsey Global Institute. 2013. *Game Changers: Five Opportunities for US Growth and Renewal.* McKinsey & Company, July.

Morrison, Wayne M. 2012. "China–U.S. Trade Issues." US Congressional Research Service report, RL33536.

Pierce, Justin R., and Peter K. Schott. 2012. "The Surprisingly Swift Decline of US Manufacturing Employment." NBER Working Paper 18655, December.

PricewaterhouseCoopers. 2012. "A Homecoming for US Manufacturing? Why a Resurgence in US Manufacturing May Be the Next Big Bet." September. http://www.pwc.com/us/en/industrial-products/publications/us-manufacturing-resurgence.jhtml.

Smith, Rebecca. 2012. "Will Truckers Ditch Diesel?" *Wall Street Journal*, May 23.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 27–48

Competing in Advanced Manufacturing: The Need for Improved Growth Models and Policies†

Gregory Tassey

In the last two decades, a growing number of emerging economies first acquired manufacturing technology from external sources and subsequently built a capability to develop it internally, thereby becoming increasingly competitive in technology-based markets. In this process of global economic convergence, Asian economies in particular have combined increasing expertise in manufacturing technology with lower labor and capital costs to "hollow out" formerly US-dominated high-tech supply chains. One result has been a steady deterioration in the US Census Bureau's "advanced technology products" trade balance (see http://www.census.gov/foreign-trade/balance/c0007.html) over the past decade, which turned negative in 2002 and continued to deteriorate to a record deficit of $100 billion in 2011, improving only slightly to a deficit of $91 billion in 2012. The bottom line is that the United States has underinvested for several decades in a set of productivity-enhancing assets necessary for the long-term health of its manufacturing sector.

The first part of this paper provides an overview of the role of advanced manufacturing as a key component of technology-based growth, explains how modern advanced manufacturing differs from traditional manufacturing, and describes the advent of global supply chains. The discussion then turns to the role of research and development in advanced manufacturing, and how it differs from the conventional simplified characterization of such investment as a two-step process in which the government supports basic research and then private firms build on that scientific base with applied research and development to produce "proprietary technologies" that

■ *Gregory Tassey is Research Fellow, Economic Policy Research Center, University of Washington, Seattle, Washington. Previously, he was Senior Economist, National Institute of Standards and Technology, Gaithersburg, Maryland. His email address is gtassey@outlook.com.*

† To access the disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.1.27 doi=10.1257/jep.28.1.27

lead directly to commercial products. Instead, the process of bringing new advanced manufacturing products to market usually consists of two additional distinct elements. One is "proof-of-concept research" to establish broad "technology platforms" that can then be used as a basis for developing actual products. The second is a technical infrastructure of "infratechnologies" that include the analytical tools and standards needed for measuring and classifying the components of the new technology; metrics and methods for determining the adequacy of the multiple performance attributes of the technology; and the interfaces among hardware and software components that must work together for a complex product to perform as specified.

For modern advanced manufacturing industries of the future, the public good content of these two additional elements implies that the process of "creative destruction" needs to be recast within a public–private investment model. Growing interdependencies exist among small and large firms, entire manufacturing supply chains, and major economic sectors (particularly high-tech services) as well as with an increasingly elaborate supporting set of research, educational, and financial infrastructures. If the public–private dynamics are not properly aligned to encourage proof-of-concept research and needed infratechnologies, then promising advances in basic science can easily fall into a "valley of death" and fail to evolve into modern advanced manufacturing technologies that are ready for the marketplace. Each major technology has a degree of uniqueness that demands government support sufficiently sophisticated to allow efficient adaptation to the needs of its particular industry, whether semiconductors, pharmaceuticals, computers, communications equipment, medical equipment, and some other technology-based industry. The conclusion of the paper explores some implications of this perspective for public policy.

Manufacturing in the Global Economy

The Shifting Role of Manufacturing in Economic Growth

The role of manufacturing in the world's economy is growing and now accounts for approximately 16 percent of global GDP and 14 percent of employment (McKinsey Global Institute 2012). In the US economy, real output in the manufacturing sector was the same in 2011 as in 1999 (Bureau of Labor Statistics n.d.). But within this sector, substantial differential growth rates have appeared among manufacturing industries. Between 2000 and 2009, the five large research and development-intensive manufacturing industries—semiconductors, communications equipment, computers, pharmaceuticals, and medical devices—had an average growth in real output of 27 percent. Meanwhile, the five large traditional industries—chemicals, machinery, electrical equipment, plastics and rubber, and fabricated metals—had an average real output growth of –23 percent (compiled from unpublished Bureau of Labor Statistics 3-digit and 4-digit real output data; see Tassey 2013a, table 1).

For a number of reasons, the advanced manufacturing industries are especially important for a future of good jobs and long-term growth. First, manufacturing provides high-paying jobs whose potential loss should not be taken lightly. In particular, many of these jobs are in research and development. National Science Foundation data show that the manufacturing sector conducts 70 percent of the research and development performed by US industry and accounts for 60 percent of industry's scientists and engineers. Allowing the domestic manufacturing sector to offshore would clearly remove a majority of the private economy's research and development capability. As the global economy spent approximately $1.4 trillion on research and development in 2013—most of it in manufacturing—continued competitiveness will require more and better research and development capabilities in that sector.

Second, on the output side, the manufacturing sector accounts for approximately 50 percent of US exports.

Third, manufacturing also generates considerable demand for support services from other sectors. For example, US-based manufacturing companies stimulate demand for 4.7 million service-sector jobs in areas such as telecom, travel, logistics, banking, and information technology infrastructure (McKinsey Global Institute 2012).

Fourth, analyses by the US Bureau of Labor Statistics show that in all but one of 71 technology-oriented occupations, the median income exceeds the median for all occupations; in 57 of these occupations, the median income is 50 percent or more above the overall industry median (Hecker 2005). Hence, the high-income economy must be the high-tech economy.

The Evolving Nature and Role of Advanced Manufacturing

For most of the post–World War II period, the typical factory consisted of standalone work stations with human-controlled machines—drilling, cutting, milling, stamping. Work passed from one station to the next by actions separated from either the sending or receiving stations. Conditions for success were dominated by the imperative to achieve economies of scale in order to meet demand for large quantities of homogeneous products at low cost.

Emerging advanced manufacturing technology is quite different. A 2011 report by the President's Council of Advisors on Science and Technology (PCAST) defines advanced manufacturing as

> . . . a family of activities that (a) depend on the use and coordination of information, automation, computation, software, sensing, and networking, and/or (b) make use of cutting edge materials and emerging capabilities enabled by the physical and biological sciences, for example nanotechnology, chemistry, and biology. This involves new ways to manufacture existing products, and especially the manufacture of new products emerging from advanced technologies.

The technologically advanced products of modern manufacturing are complex systems. Automobiles used to be largely a modestly complex set of hardware components: engine, drive train, suspension, and the like. Today, the modern automobile contains 17 subsystems for which electronics is a central element. These subsystems are controlled and connected to each other by nearly 100 microprocessors and five miles of wiring (Kurfess 2011).

The production system that manufactures the modern automobile and other complex system technologies is increasingly based on computer-controlled and integrated systems of connected stations, which generate and analyze data on multiple attributes of products at multiple phases of completion, thereby allowing real-time adjustments. Even the transfer of the product between work stations is automated. As the demand for customized or at least semi-customized products continues to grow, manufacturing processes must be increasingly flexible to achieve the economies of *scope* required to serve a set of heterogeneous submarkets. Flexibility while still maintaining high performance and low unit cost can only be achieved through new processing techniques, ubiquitous use of information technology, and a highly skilled labor force. Finally, industry structure will be significantly altered as new production concepts such as 3D printing will enable a range of products to be produced quickly at one installation, thereby ushering in an era of what is coming to be called "manufacturing as a service." So radical are these characteristics of advanced manufacturing and its projected economic impacts that some analysts have labeled this emerging technology as the "fourth industrial revolution."

For this to happen, the above set of manufacturing technologies will have to evolve in terms of both physical characteristics and the complex set of information technology–driven processes that create them. Clearly, part of this paradigm shift is increased automation. Thus, we are witnessing a replay of historical fears of the impacts of automation on employment that have occurred periodically for more than a century (for example, Krugman 2013). However, if automation does not occur, then a domestic manufacturing industry will either move off-shore to low-cost labor economies or, increasingly, to advanced economies that have made the requisite investments in advanced manufacturing technologies. If automation does occur, the labor content per unit of output will decline, but the increased productivity will result in larger global market shares and hence an increased demand for well-paid labor.

One sometimes hears the view that with a combination of automation and outsourcing, advanced economies can become basically service economies, doing cutting-edge scientific and design work together with marketing, distribution, and retail, while the actual manufactured goods are imported from other countries. However, this compartmentalized view of the modern economy is inaccurate. As described in the following section, advanced manufacturing often displays important co-location synergies resulting in benefits to new-product development when manufacturing firms are located close to their research and development efforts and to many of their key suppliers. These synergies arise from the fact that much of the technical knowledge developed in the early phases of the research and development cycle is tacit in nature (as opposed to being codified in, say, patents). As a result, person-to-person interactions are critical to advancing and transferring such knowledge (Goffin and Koners 2011).

In addition, technological complexity demands close interaction between high-tech manufacturing and the service systems (finance, medical, transportation, communication, and so on) that depend on the manufacturing sector's technologies. As a result, many larger R&D-intensive companies with a manufacturing core have substantial (largely business-to-business) service units in an attempt to capture synergies between these two sectors of economic activity. For example, as part of a companywide restructuring in the 1990s, IBM created a strategic focus on software services. In 2012, the company derived 41 percent of its revenues from such services. Similarly, Hewlett-Packard purchased Electronic Data Systems in 1999 as a major strategic step to integrate forward into services, and in 2012 services accounted for 35 percent of revenues. McKinsey Global Institute (2012) estimates that across manufacturing subsectors, service-type activities account for 30 to 55 percent of employment.

Supply Chains for Advanced Manufacturing

Standard microeconomic theory emphasizes decisions of firms and the interaction of firms within the same industry. However, the growing complexity of manufacturing technology has increased the distribution of production and of research and development among a series of related industries. An approach that looks at firms or at industries one-by-one gives insufficient attention to the role and interactions among these industries in a modern technology-based economy.

Co-Location Synergies and the Modern High-Tech Supply Chain

The supply chain is the key unit of analysis for understanding these interdependencies. When domestic consumer-product manufacturers lose market shares, the domestic supplier industries that support them tend to contract as well. As one example, the increasingly sophisticated machine tool industry, which is essential to all discrete parts manufacturing industries, was once a highly competitive US industry. But the US share of global output in machine tools has declined from 20.4 percent in 1980, to 9.8 in 2000, and to 5.3 percent in 2012, according to Gardner Business Media's *The World Machine Tool Output and Consumption Survey*. A major reason is that manufacturing industries using machine tools are increasingly located in other countries, which creates incentives to produce these tools in those economies, near their business customers. China became the largest consumer of machine tools in 2002 and a decade later consumed more than four times as much in machine tools as the US economy. Not coincidentally, China now produces more than five times the US output of machine tools.

The companies at the ends of manufacturing supply chains have been referred to as "original equipment manufacturers," meaning such companies actually produce the product—even if that product is later sold under a different brand name. In most traditional supply chains, the original equipment manufacturer has been the dominant firm in terms of technology development. In the US automobile industry, for example, the "Big Three" of General Motors, Ford, and Chrysler traditionally did most of the research and development for automobiles, and then sent component specifications back up the supply chain for production of these components and even subsystems. However, as the technology content of modern manufactured "systems" (autos, airplanes, personal computers, smartphones, and many other advanced manufacturing products) increased, original equipment manufacturers were forced to outsource increasingly larger percentages of their system technology needs. In effect, research and development migrated backward to earlier tiers (industries) in these supply chains, a phenomenon that has led to greater challenges for managing supply-chain coordination.

A common dynamic is that integrated manufacturers dominate the supply chain for a period of time until the interfaces between components are firmly established and the markets become large. At that point, standardized interfaces and economies of scale allow innovative specialists in individual components to enter the industry. The tier in a supply chain at which this vertical disintegration occurs, known as the "decoupling point" (Christensen, Musso, and Anthony 2004), tends to move backward in the supply chain over time from the final product to subsystems and then to component tiers. As a result, the level and complexity of interactions among these tiers increases over a technology's lifecycle.

By integrating electronics industries, Asian economies have captured co-location synergies and taken over increasingly large shares of the value added in electronics products supply chains. In apparent contrast, many American firms have adopted "design-only" strategies. As one prominent example, Apple does some prototype manufacturing in the United States, but off-shores all component manufacturing and assembly. Samsung, its leading competitor in smartphones, has the advantage of being a co-located vertically integrated electronics company, as evidenced by its ability to turn out new versions of its smartphones at a faster pace than does Apple. Further, the overall quality of its phones has increased rapidly. Although Apple was the innovator in smartphones, Samsung has a 31 percent share of the global market, compared to 13 percent for Apple (AFP 2013).

Still, a superior domestic innovation infrastructure and the difficulties experienced by competing economies in imitating can keep innovative activity within the first-mover economy—at least for a while. An example is semiconductors where manufacturing has moved offshore to a substantial degree, but indicators of inventive activity (as measured by relative patent rates) still show a US-centric pattern, especially for process technologies (Macher, Mowery, and Di Minin 2008). "Fabless" (no manufacturing) semiconductor companies have been successful in the current mature phases of the current CMOS technology lifecycle by adopting highly accurate simulation techniques that drastically reduce the number of expensive and time-consuming iterations of a product's design necessary to enable its manufacture. Off-shore dedicated foundries (the producers), in turn, often do not even operate development-scale fabrication facilities, instead relying on real-time adjustments. Both of these single-phase strategies can work within the middle and later phases of a particular technology's lifecycle. But when a new technology lifecycle begins, both strategies may hit a brick wall. The “fabless” firms will not be able to execute design for new manufacturing requirements without close interaction with manufacturing scale-up activity, and foundries will not be able to adapt to radically new product technologies without close interactions with the ongoing product research and development. Thus, the process of off-shoring the manufacturing component of high-tech supply chains has long-term negative consequences, and especially so, the greater the complexity of the technology and the earlier the technology is in its lifecycle.

A number of Asian economies—particularly China, Taiwan, and Korea—have followed growth strategies that started with a few high-tech industries for which major efforts were made to become competitive. These economies aggressively invested in the five major asset categories characterizing technology-based competition: intellectual capital, physical capital, human capital, industry structure, and technical infrastructure. Such investments have enabled these economies to integrate forward (Korea, Taiwan) or backward (China) from the entry industry (Tassey 2010).

Therefore, global competition is no longer just about the success of individual firms or even individual industries. Instead, it is about achieving and maintaining a competitive position in a global supply chain. The current process of global convergence among the world’s economies is a repeat (although faster version) of a phenomenon that has occurred a number of times over the past several centuries (Lucas 2009). However, there is no intrinsic reason why leading economies have to be caught and passed by determined upstarts—either within or between technology lifecycles. In fact, in the future, no one economy will likely dominate an entire supply chain. Global technology-based competition has simply become too vast. Thus, “second best” strategies will likely be the compromise objective: an economy will attempt to be one of several successful competitors in at least several tiers of a high-tech supply chain.

Supply-Chain Benefits and Costs

The economic rationale for dispersed global supply chains is based on the benefits of decentralized low-cost production and technology specialization. This process is facilitated by ever-improving information and computer technology and the decreasing weight/value ratios of many high-tech products (so that transportation costs are less of a barrier to geographic dispersion). Research and development networks spread risk and combine complementary research assets through a process called “open innovation” (Chesbrough 2004). Such networks, using advanced information and communications technology, have become an integral part of global supply chains and can often function quite well. The underlying premise is that these gains in cost and asset allocation efficiency can overcome language and cultural barriers, currency swings, differences in intellectual property laws, and reduced person-to-person interactions between supplier and customer.

But over longer periods, spreading the supply chain geographically reduces the speed and usefulness of information transfers made possible by co-location, especially during and immediately after transitions to new technologies. In particular, during the early phases of the research and development cycle and again during production scale-up after initial commercialization, new technical knowledge is unsettled, thereby requiring constant adjustment. The tacit nature of such knowledge and the frequent need for its transfer is facilitated by continual person-to-person exchanges.

With this potential tradeoff in mind, advanced manufacturing firms considering outsourcing some key components must weigh short-term production cost reductions for a current technology against the risk of falling behind in the next technology cycle. For example, Boeing's move to a "value-stream" supply chain strategy in which suppliers are located in many different countries created coordination problems that seem to be largely responsible for the two-year delay in flight testing the 787 Dreamliner (Petrick 2009). The geographic dispersion of suppliers and Boeing's consequent difficulties in restructuring its supply chain strategy greatly increased management costs, as well as time delays, and Boeing eventually had to bring some of the subassembly work back in-house (Tang and Zimmerman 2009).

The US printed circuit board industry was once relatively labor-intensive, which led to its off-shoring. Today, its production process is highly automated, with low unit labor content, but the transition to automated production happened in other countries where downstream industries are located. Thus, the majority of the global industry remains in those locations (Asia) near the subsequent tiers in the electronics supply chain—component and final product assembly. According to IPC-Association for Connecting Electronic Industries, the US accounts for less than 5 percent of world production, while Asia's share is 90 percent (IPC 2012).

An example of a current technology lifecycle transition is optoelectronics, which is an increasingly important industry because virtually all of the data moving across the Internet is based on optical systems (photons generated by laser diodes replacing electrons generated by transistors). But the current technology lifecycle is largely exhausted.

Thus, optical circuit technology is beginning to undergo a technology lifecycle transition. The current version of optoelectronics is a discrete technology; that is, each component has its own substrate/chip/package, so that individual components must be physically connected to form a more complex device. However, the emerging new generation of optoelectronics will be an integrated technology; that is, multiple components are simultaneously created on the same substrate, which reduces overall size and improves functionality (faster, smaller, less heat produced). The mature *discrete* technology can be produced more cheaply in Asia. However, if the discrete technology is produced in Asia, then the "installed base" of private economic assets and supporting public infrastructure increase the likelihood that the integrated technology will be established there, too (Fuchs, Field, Roth, and Kirchain 2011). If US firms are to become innovation leaders in the new *monolithic* integrated technology, public and private investment strategies will be needed to accelerate the evolution of and scale-up for initial markets, in spite of the fact that they would be at a cost disadvantage compared to discrete production for a period of time—a common situation when a new potentially superior technology is initially attempting to penetrate a market (Tassey 2013b).

In short, while the potential for short-term gains from globalization of supply chains can be clear-cut for individual companies in the middle and later portions of a technology's lifecycle, the loss of co-location synergies in domestic supply chains and the degradation of the domestic economy's private and public research and production infrastructures make being competitive in the next technology lifecycle more difficult. Repeating this scenario across advanced manufacturing sectors will lower the long-term rate of domestic economic growth.

Metrics of Research and Development Policy: Quantity, Composition, and Efficiency

Researchers would benefit from a better characterization of the research and development process that produce modern industrial technologies. To this end, research and development can be divided into three major dimensions: amount, composition, and efficiency.

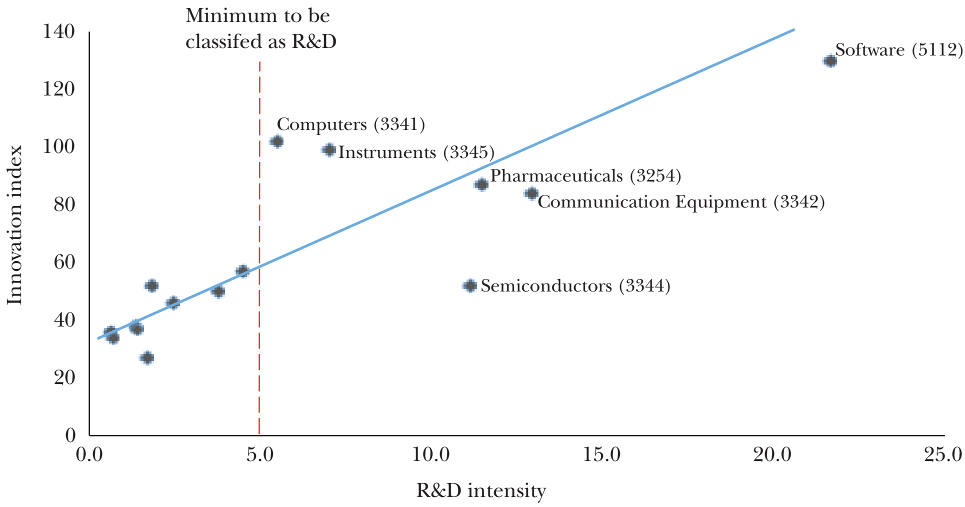
The Quantity of Research and Development Spending

The importance of the quantity of investment in development of new technologies with respect to its impact on innovation can be indicated using data on product and process innovations recently compiled by the National Science Foundation for a broad cross-section of industries. The horizontal axis of Figure 1 shows the research and development intensities for 4-digit NAICS industries or 3-digit NAICS industry groups as measured by the amount of research and development spending by an industry divided by net sales (average for 2003-2007). On the vertical axis, a rough index of industry-level innovation is created by adding the shares of companies reporting product and process innovations for each industry over the period 2006–2008. Product and process innovations are reported separately, so the maximum possible value on the vertical axis would be 200 (percent).

The vertical dashed line in Figure 1 indicates the 5 percent ratio of research and development to sales that has commonly been used (by the OECD, for example) to classify an industry as research and development "intensive." In the figure, 9 of the 14 manufacturing industries fall below this minimum (software is also included because of its ubiquitous use within the manufacturing enterprise, much of it embodied in physical products). For the sector as a whole, NSF data show that US manufacturing's average research and development intensity has increased over a 25-year period from 2.6 percent in the early 1980s to 3.7 percent in 2007. When government funding of US manufacturing company research and development is added, the R&D intensity for 2007 increased modestly to 4.1 percent.

To some extent, any economy will have a mixture of industries that are more-or-less research and development intensive. But the level of research and

Figure 1
Rate of Innovation versus R&D Intensity in US Manufacturing



Sources: Data using data from *Science and Engineering Indicators 2010*, Appendix Table 4-14 for R&D intensity data and Boroush (2010) for innovation data.
Notes: The horizontal axis of Figure 1 shows the research and development intensities for 4-digit NAICS industries or 3-digit NAICS industry groups as measured by the amount of research and development spending by an industry divided by net sales (average for 2003–2007). On the vertical axis, a rough index of industry-level innovation is created by adding the shares (in percent) of companies reporting product and process innovations for each industry over the period 2006–2008. Product and process innovations are reported separately, so the maximum possible value on the vertical axis would be 200 (percent). Note that in each of two pairs of industries (extreme left in plot area), the industry markers overlap.

development should not be taken as an intrinsic limit reflecting innovation potential. After all, technology-based economic growth can include revitalizing existing industries that are currently "low-tech," requiring a higher research and development intensity.

For the economy as a whole, economists often argue that private firms will underinvest in research and development relative to the social optimum. Indeed, there are six characteristics of research and development that can limit the willingness of advanced manufacturing firms to make appropriate levels and types of investment: 1) high technical and/or market risk with respect to whether the research and development investment will pay off; 2) lead times (investment to commercialization) that are typically longer than for other types of investments; 3) knowledge spillovers of two types: a) the intrinsic nature of knowledge that allows it to "leak" to competitors and b) multiple potential product markets emanating from a technology platform (economies of scope), some of which are beyond the market focus of the investing firm, resulting in a lower expected rate of return; 4) price spillovers in which an innovative firm has insufficient pricing power to capture monopoly rent; 5) information asymmetries between buyers and sellers of technology-based products (buyers cannot verify performance of new products resulting in excess transaction costs and thereby slower market penetration); and, 6) coordination failures between adjacent industries in supply chains that reduce the potential for "open innovation," that is, cooperation in research and development.

Economic studies like Jones and Williams (1998, 2000) have estimated the return on research and development to be four times the return on investment in physical capital, implying that it should increase dramatically. But for modern emerging technologies, the scope of the overall required research on certain key topics is increasingly beyond the research capabilities of even large firms. For decades IBM, like other large research and development-intensive companies, conducted virtually all research within its corporate boundaries. Today, in order to be competitive in the emerging field of nanoelectronics, IBM is a major investor and participant in the Institute for Nanoelectronics Discovery and Exploration located at Albany State University. IBM is partnering with other electronics companies (like Intel, Micron, AMD, Texas Instruments, and Freescale Semiconductor, Inc., as well as with smaller companies who are often suppliers to IBM) and with several universities. The New York State government is a major financial supporter, and the federal government is directly involved in the research (Welser 2008). This investment would be unlikely to happen in anything like the same scope and quantity without this kind of extensive public–private cooperation.

The Composition of Research and Development Spending

US technology policy has traditionally been built on a "black-box" model in which government primarily supports basic science, a consensus public good. In this model, science is then turned into innovations through a private investment sequence based on the view that "technology" is a pure private good. However, corporate organization and investment behavior clearly show that technology investment is not homogenous and that technologies are not black boxes. Rather, the transition from basic science to commercial product includes three major elements: technology platforms, infratechnologies, and proprietary technologies. The two additional elements, technology platforms and infratechnologies, exhibit degrees of public good content.

Once a science base is in place, "proof-of-concept" technology research (as the management literature calls it) typically occurs long before commercialization. The result of such research is the creation of a broad technology-platform, whose existence both confirms the potential for multiple market applications and provides a set of technical conceptualizations that drive applied research and development. One of the best-known examples of proof-of-concept research is Bell Labs' demonstration of the technical concept that semiconductor materials can be organized to perform the functions of an electronic switch or amplifier. The existence of broad economies of scope in proof-of-concept research, together with the high degree of technical and market uncertainty, means that individual firms would be unlikely to capture many of the potential markets from doing this type of research.

To illustrate the problem, the proof-of-concept phase of research and development in the pharmaceutical industry would involve a full conceptual model of the proposed new drug mechanism: specific biological targets, bioavailability, toxicity, and the like. For several decades, the National Institutes of Health spent billions of dollars on life-science research and then waited for private venture capital to fund the development of new proprietary biopharmaceuticals. However, private risk capital does not like to invest in proof-of-concept research—due to the long time to market and the remaining high technical and market risk. So biopharmaceutical firms have attempted to develop new drugs directly from the underlying science, basically by using clinical trials (specifically, Phase II trials) to prove the new drug concept. This approach largely skips a true proof-of-concept phase, because a drug candidate needs to be formulated before a clinical trial, whose purpose therefore is to initially test a specific application of a presumed technology platform. If the technology platform (concept) is not fully developed, application efforts are more likely to be unsuccessful. In fact, Gallaher, Petrusa, O'Conner, and Houghton (2007a) show that Phase II trials have exhibited low probabilities of success for drug candidates advancing through additional clinical testing to approval by the Food and Drug Administration.

Infratechnologies are a diverse set of technical tools that are necessary to conduct all phases of research and development, to control production processes, and to execute marketplace transactions for complex technology-based goods. They include research tools like measurement and test methods, scientific and engineering data, quality control techniques, and the functional as well as physical basis for the interfaces between components of modern technology systems. These tools are called "infratechnologies" because they provide a complex but essential technical infrastructure, which is as critical to achieving adequate private investment for the modern technology-based economy as traditional economic infrastructure was for the Industrial Revolution.

Infratechnologies are often embodied in the standards that are ubiquitous in high-tech industries. The semiconductor industry has over 1,000 standards without which that industry could not function. Gallaher et al. (2007b) estimated that this industry spent $12 billion on measurement infratechnologies in the period 1996–2006, which generated gross benefits of $52 billion in 2006 dollars. Without the availability of this technical infrastructure (most of which is codified as industry standards), costs would be higher not just at the research and development stage, but also during production and even marketing.

In summary, proof-of-concept research, infratechnologies and applied research and development exhibit different degrees of public good content and hence distinctly different investment incentives. Each of these technology elements therefore requires a unique set of policy responses. Basic science is close to a pure public good, which is why it makes sense that the lion's share of basic research is funded by government. Proof-of-concept (technology-platform) research and infratechnology research combine public and private good aspects and are typically co-funded by industry and government; so there is an overall rationale for the evolving partnership mechanisms increasingly observed in the global manufacturing economy. The third element, proprietary technology, is closest to a pure private good, but even in this case, relatively high risk leads to underinvestment, which explains the existence of a "research and experimentation" tax credit.

Clearly, an updated knowledge production function is needed that can better address the major technology elements and their interactions among each other and with the supporting technical infrastructure. In Tassey (2007, 2010), I offer a disaggregated three-element technology production function—showing how the science base is drawn upon to create technology platforms supported by infratechnologies, which then combine to enable the "proprietary technologies."

The Efficiency of Research and Development Spending

The efficiency of research and development policy refers to the organization of research efforts relative to the structure that optimizes the return-on-investment. The appropriate organization will consider the portfolio of technologies needed to achieve complete development of system technologies, with due attention paid to the relative amounts of investment required for each component of that system. It will consider especially the distribution of research and development funding by technology element across the phases of the research cycle: basic research, proof-of-concept research, and applied research and development (infratechnologies are developed and used in all phases). It will further optimize the mix of participants (universities, government, and industry), the mechanisms by which public and private actors collaborate (ecosystem attributes), and the roles of research and development infrastructure (like research facilities and skills of researchers).

Indeed, the attributes of research and development efficiency require a complex organizational format, which is rapidly evolving among the world's technology-based economies. The single most important emerging format, the "regional innovation cluster," has become a global phenomenon. This organizational innovation is a more elaborate version of the stand-alone research consortia that began to appear in the 1980s. The best known of these is SEMATECH, which facilitates development of semiconductors and related production equipment technologies. The cluster model increases the efficiency of technology-based economic growth strategies through co-location of public and private research and development assets within a local area or region. Co-location synergies are achieved through use of the research consortium mechanism, a supporting education infrastructure, involvement of multiple industries in the emerging supply chain, and shared scale-up production facilities to accelerate commercialization.

Clusters can also provide concentrated labor pools with the relevant skills, as well as promoting technology diffusion and hence broader commercialization of research results. A fully functioning innovation cluster can enable management by the entire supply chain of successive technology lifecycles. Moreover, the research consortium facilitates effective management and sharing of intellectual property.

"Additive manufacturing"—frequently called 3D printing—is an example of a complex advanced manufacturing technology that can benefit from more efficient research mechanisms, such as the research cluster. Traditional manufacturing techniques take bulk materials and then drill, cut, mill, or stamp—usually wasting considerable material in the process and ending up with a part made from a homogeneous material. In contrast, additive processing techniques build products by sequentially adding layers of different materials in different combinations and configurations. This approach allows much more complex and varied products or components to be produced by a single production unit, dramatically reducing the retooling and assembly steps. As described by Manufacturing.gov (at http://www.manufacturing.gov/nnmi_pilot_institute.html), "key benefits of additive manufacturing are that it enables shorter lead times, mass customization, reduced parts count, more complex shapes, parts on demand, less material waste, and lower life cycle energy use."

Additive processing technologies require sophisticated materials management and assembly techniques, which must be developed by multidisciplinary teams who first must prove the overall technical concept before processing steps can be specified and integrated into a 3D printing device. Such proof-of-concept research and the follow-on applied research and development are measurement intensive and require detailed processing data and computer models. For example, the average thickness of a layer of material deposited by a 3D printer is now less than 100 microns (a micron is one millionth of a meter). Without sophisticated measurement infratechnologies, the required technology platforms and subsequent product innovations would not be realized.

In summary, thinking about research and development as a heterogeneous, multi-asset investment process, rather than as a two-dimensional mixture of basic research and applied research and development, suggests that the overall efficiency of an economy's research and development effort will be based not just on the level of spending, but on the composition of that spending and how that spending is managed.

The Shifting Strategies of Advanced Manufacturing Firms

When the United States was the world's dominant technological power, large US companies could apply lower discount rates to longer-term, higher-risk research and development projects because they faced relatively little competition and therefore longer technology lifecycles. Technological complexity was sufficiently low so that a single large company could have most of the research and development assets required to pursue major breakthroughs. The provision of infratechnologies and associated standards often evolved slowly over a technology's lifecycle due to inadequate research funding and inefficient standard-setting institutions. However, in the face of weak foreign competition, such lags had minimal negative effect.

The absence of significant competition also allowed companies to expect that they would achieve economies of scope over time by capturing multiple markets based on new technology platforms. For example, IBM once produced an array of electronic components for retail markets in addition to its main product focus, computers. Eventually, it became too difficult to compete with Intel and other specialized chip companies, so IBM shifted to a narrower focus on specialized semiconductor components for use in its own devices. It eventually gave up on the PC, which it helped pioneer. Hewlett-Packard was also both a vertically and horizontally integrated company for most of its history, but in the last 20 years it left multiple lines of business, including measurement and testing equipment for which is was a market leader, retaining only computers, storage, and imaging (printers). Even within these categories, HP narrowed its product scope. For example, in 2011, the company withdrew from smartphones and tablet computer markets, while maintaining its personal computer business line.

For decades, manufacturing corporations have developed technological proof-of-concept (technology platforms) in their central corporate research labs. The applied research and development that results in proprietary technologies is then assigned to the research facilities in the company's line-of-business units. Many companies also have dedicated laboratories to assimilate/develop infratechnologies and associated standards—for example, analytical laboratories in chemical and biopharmaceutical companies, and metrology labs (labs related to the science and technology of measurement) in semiconductor companies.

More recently, high-tech companies have been allocating more of their research and development budgets to short-term applied research and development, aimed at maximizing profits in the middle and later phases of technology lifecycles, at the expense of long-term research aimed at new technologies (according to 20 years of firm survey data from the Industrial Research Institute discussed in Tassey 2013a, figure 5). Thus, their central research laboratories are receiving a declining share of corporate research and development funds, and increasing portions of these laboratories' budgets are allocated to supporting their business units' applied research and development programs or to assessing external sources of new technologies.

It seems likely that this reallocation of investment is also adversely affecting infratechnologies, which are themselves increasingly complex and often derive from a different science base than the industry's core technology. To understand the role of the public–private good nature of infratechnologies, the National Institute of Standards and Technology (n.d) conducted approximately 15 prospective economic studies of the costs of inadequate infratechnologies across a wide range of technologies and industries, using industry survey data. The results indicate significant underinvestment by industry due to the public good characteristics of the infratechnologies. A larger number of retrospective economic impact studies were conducted of specific government infratechnology research programs based on industry survey data. The vast majority of these studies yielded high estimated rates of returns (Link and Scott 2011). In the few studies where the net economic benefits from government infratechnology investments were low, the analysis indicated that the infratechnology research was being conducted to a sufficient extent by industry. Thus, for the most part, the government infratechnology research programs were found to be addressing significant existing market failures.

The systematic underinvestment by industry in these two quasi–public good technology elements implies that the way in which modern industries seek to address proof-of-concept and infratechnology issues is closely bound up with the ways in which government supports that industry. The pharmaceutical industry has basically tried to force its way to innovation through continued application of the black-box model—in which individual companies seek to build directly on basic science results without sufficient support for proof-of-concept or infratechnology research—with limited success. In contrast, the semiconductor industry has embraced research consortia and investment in infratechnologies and standards. Of course, these industries differ in other ways, too, but part of the difference in their ability to innovate in recent decades can be explained by differences in the type of government support.

To address these underinvestment patterns, industrialized nations are moving at various rates toward the technology-element model described above. Location of investment by global high-tech companies is increasingly affected by the relative efficacy of national investments in the public good portions of these elements. With respect to the previous example of additive manufacturing, the Obama administration has established a National Network for Manufacturing Innovation. One of the first implementations is the National Additive Manufacturing Innovation Institute (NAMII) in Youngstown, Ohio. Importantly, this location is in an emerging technology cluster in the Eastern Ohio–Western Pennsylvania area. As one indication of how such institutional strategies can even attract foreign investment, the giant Asian electronics manufacturer, Foxconn, a "notorious low-wage manufacturer of Apple's iPhone" and hence "a poster child of U.S. outsourcing," will invest $30 million in a new robotics plant in Harrisburg, Pennsylvania. Foxconn was attracted by major elements of the cluster: the NAMII and Carnegie Mellon's world class robotics program to which Foxconn has contributed another $10 million (Muro and Andes 2013).

An Updated Growth Model for Advanced Manufacturing

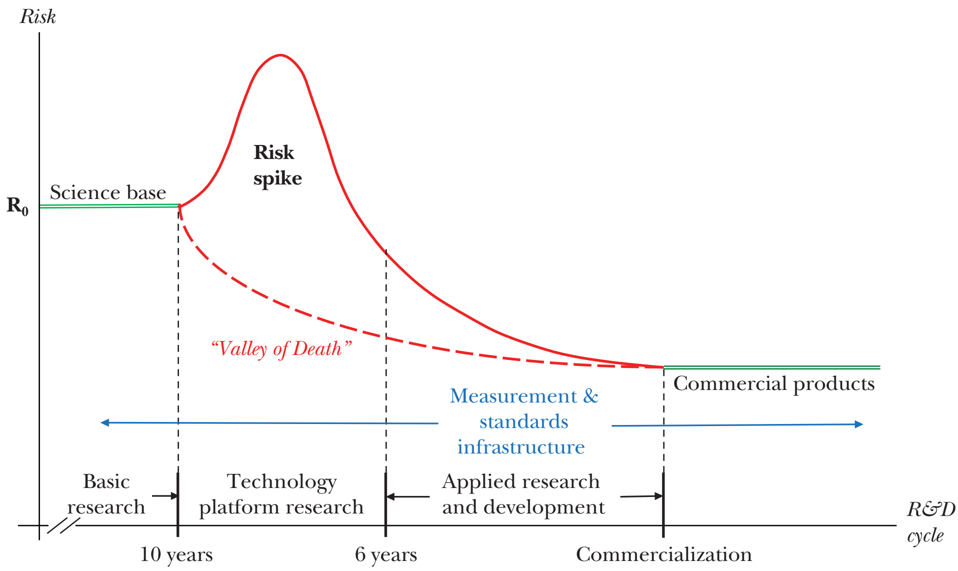
Robert Solow (1956, 1957) characterized the role of technology by using a production function to estimate how much of economic growth could be explained by increases in capital and labor, and then attributing the residual to technology. In these early estimates, the technology residual accounted for about 40 percent of US economic growth. However, both the nature of technology as an economic asset and the investment process by which technical knowledge is created remained to be specified. Moreover, in the traditional economic growth models, technology is a pure private good derived from some exogenous source, which is an implausible assumption for discussions of modern technology policy. Further, the introduction of technology changes the relative productivities of all inputs, as well as the product structure and therefore drives the optimal capital/labor ratio.

This fact alters comparative advantages and creates a second role for technology, that of enabling "adaptive efficiency" as an economy evolves and grows in a global economy (Audretsch and Link 2012). Indeed, the pervasive role of technology has altered the concept of comparative advantage as a superior level of efficiency for one country due to a unique and stable set of endowed assets that create the efficiency advantage. Long ago, a comparative advantage once established could be expected to last for some time. After all, countries' endowments of relatively fixed assets (arable land, mineral deposits, navigable waterways, and climate) remained relatively unchanged and therefore so did relative prices. As manufacturing became a larger share of industrializing economies, the relative efficiencies of deploying capital and labor became more important, hence the focus in Solow's growth model on optimal capital/labor ratios. The optimal capital/labor ratio was determined in part by an underlying set of technologies. However, until the last 50 years, technology changed sufficiently slowly and its sources were frequently considered exogenous so that its role was not explicitly included in growth models. But in the modern economy, the ability of technology to change comparative advantage on a relatively frequent basis makes this input an active part of corporate and national growth strategies, rather than a passive "initial condition."

Three decades after Solow's early work, Romer (1990) introduced an explicit knowledge production function into a general equilibrium growth model. In defining knowledge, Romer distinguished between "rival" and "nonrival" elements of technology. The rival (or excludable) component of technology was embodied in human production capital (and thus part of the tacit component of knowledge referred to earlier). The nonrival component—which in Romer's conceptualization seems to be a disembodied technology element not linked to any specific factor of economic activity—is viewed as partially excludable, and partially not. In microeconomic studies, the production of technical knowledge was mostly characterized as resulting simply from a distribution of past research and development expenditures. Only a very few economists have indicated the possibility of more than a single technology element (Mansfield 1980; Griliches 1986; David and Hall 2000).

Modeling a multi-element technology production function, as described conceptually in the previous section, poses some analytical challenges. It involves specifying multiple categories of technical knowledge, directions of technology element flows and their interaction mechanisms, and degrees of excludability for the different kinds of technology. Within an industry, spillovers occur from one firm to another, among a consortium of firms acting collectively, or to all firms from a source exogenous to the industry (frequently from government or universities). Further, the elements in a technology production function interact with each other. For example, the technology platform drives the productivity of applied research and development. Infratechnologies leverage the productivity of all phases of research and development and also production. For more detail and a formal development of this kind of production function, see Tassey (2005).

Figure 2
Technology Platforms and the Research and Development (R&D) Risk Spike



Source: Tassey (2008).

A critical concept underlying a technology's development and eventual commercialization is a process of risk reduction, as indicated in Figure 2. The horizontal axis is time. The vertical axis represents the level of risk faced by a private-sector technology-based firm considering a course of action that can lead to a commercial product. The traditional view of research and development assumes a smooth pattern starting with a base of scientific knowledge and developing progressively more applied knowledge until a commercial product is produced, which is a pattern of monotonically declining risk indicated by the dashed line.

But when industry begins to consider the possibility of developing commercially viable technologies from an existing science base, it must take into account the significant technical and market risks related to potential commercialization. This significant investment barrier or "risk spike" must be overcome before substantial private investment in applied research and development will be forthcoming. The risk spike is the so-called "valley of death" referred to in innovation policy circles. At this point, the platform technology is immature and the appropriate infratechnologies are ill-developed; thus, initial attempts at innovation solely through company-funded applied research and development frequently fail.

Indeed, on average, the greater the potential of a new technology, the greater the required advance in technology platform development and infratechnologies. That is, the risk spike in Figure 2 will typically be larger for a radically new technology (like one based on a recent major scientific breakthrough) than for investment in less radically new technologies (the next generation of an existing platform technology). Such a risk profile explains why rates of technical progress in the early phases of the research and development cycle targeting radically new technologies can languish for years. However, once the risk spike is overcome, private investment in research and development can flow at sufficient rates to achieve commercialization.

The Policy Imperative

The vast majority of economic debate in recent years has been over ways to overcome the Great Recession and its aftereffects. But policies for business cycle stabilization (monetary and fiscal tools) are far from a holistic long-term growth strategy (Tassey 2013a). In contrast, this paper focuses on structural issues and hence on the strategies that must be resolved for US manufacturing to be globally competitive in the long run. The model of industrial technology development described here implies four important modifications to technology-based economic growth policy.

First, innovation clusters need to be embraced and actively supported. The complexity of modern manufacturing technologies and the shrinking "windows of opportunities" due to intense global competition demand not only more and better-balanced research and development investment, but also more efficient research infrastructures. Specifically, this means geographically concentrating research and development assets from multiple sources and involving several tiers from the emerging high-tech supply chain.

Second, finding ways to reallocate labor as comparative advantage shifts is becoming more important. A few economies, notably Germany, are adept at "reuse of legacy capabilities"; that is, adapting processing and other skills developed for one industry to a newer one (Berger and MIT Industrial Performance Center 2005). This adaptive strategy with respect to labor greatly expands and sustains the acquired expertise in specific processing technologies. The Japanese were also particularly good at "gijutsu yugo" or "technology fusion" during their economy's period of high growth in the 1980s when, for example, they transferred semiconductor processing technology to the new area of optoelectronics (Tassey 1992).

Third, smaller firms are suffering to a greater degree from increasing foreign competition (Petrick 2009). The American Small Manufacturers Coalition (2009) estimates that one-third of small manufacturers (90,000 firms with sales less than $10 million in annual revenue) are not at or near world-class in any element of corporate strategy. For larger firms (more than $100 million in revenue), a smaller share (14 percent) are estimated to be equally deficient. As pointed out by Audretsch and Link (2012), Schumpeterian "creative destruction" initially emphasized the role of the entrepreneur, and hence small firms, as the engine of innovation. This focus resulted from the need for small firms to find a way to disrupt the established markets dominated by large firms. Schumpeter eventually reversed his view and emphasized the superior capabilities and market strength of large firms, which enabled them to be more efficient and successful innovators. Today both large and small firms coexist in the same technology-intensive supply chain. Each provides complementary assets in the form of components (hardware and software) and their integration to yield the final technology system. In the worldwide emergence of innovation clusters, firms of all sizes agglomerate into integrated supply chains that deliver emerging technologies through a highly distributed pattern of research and development.

Fourth, government policies regarding research and development and innovation will play an important role given the importance of the public–private nexus as advanced manufacturing firms seek to shepherd new technologies through the valley of death to become commercial products. To this end, the current trends are troubling. Federal research and development funding as a percent of GDP was 54 percent lower in 2011 than in 1964, the peak year for this intensity ratio (National Science Foundation n.d.). Moreover, it is not just a matter of the amount a government and the economy as a whole spend on research and development but how it is spent—across industries, over phases of the research and development cycle, among tiers in high-tech supply chains, and through various research infrastructures. The mechanisms for encouraging research and development go beyond direct spending, and also include encouraging various forms of risk-sharing and cooperation not only among competing firms with respect to development of quasi–public-good technology elements, but also between government and industry.

From a broader perspective, the US advanced manufacturing sector faces a potential long-term inadequacy of investments in the five categories of productivity-enhancing assets: 1) technology and intellectual capital; 2) skilled labor; 3) hardware and software; 4) industry structure and behavior (organizational/marketing capital); and 5) technical infrastructure. This inadequacy is not appreciated. Instead, one hears manifestations of an "installed wisdom" effect, which refers to the almost endless stream of rationalizations of how the US economy will continue to be highly competitive without significant change (Atkinson and Ezell 2012, chap. 4, provide an excellent set of examples).

The array of emerging advanced manufacturing platforms will shift leadership positions among the world's economies. No single economy has yet implemented a complete model for creating and managing a series of domestically focused technology lifecycles. However, some nations like China and Korea in Asia and a number of northern European economies are embracing investments in the required productivity-enhancing assets to a greater degree than others. A few northern European economies have established competitive manufacturing sectors that attain regular trade surpluses, in spite of the fact that average hourly labor compensation costs are much higher than those in the United States (see Bureau of Labor Statistics 2012, http://www.bls.gov/opub/ted/2012/ted_20121221.htm).

Thus, the future of US advanced manufacturing will be determined not only by the efforts of individual companies, although such efforts are of course indispensable, but also by the extent to which the US public–private system for bringing new waves of technology to market is updated and reformed. In the modern global economy, national governments compete against each other as much as do private firms.

■ *The author gratefully acknowledges comments from David Autor, Chang-Tai Hseih, and Timothy Taylor. This paper was reviewed and approved by an internal review board of NIST in February 2013.*

References

AFP. 2013. "Samsung Extends Smartphone Lead over Apple." October 29. http://www.google.com/hostednews/afp/article/ALeqM5ik_noP2mAJhhjanFWYkShVR6qQCQ?docId=c1f3cbed-c8a1-4321-98a2-6782a9141db0.

American Small Manufacturers Coalition. 2009. *Next Generation Manufacturing Study.* (June)

Atkinson, Robert D., and Stephen J. Ezell. 2012. *Innovation Economics: The Race for Global Advantage.* New Haven: Yale University Press.

Audretsch, David B., and Albert N. Link. 2012. "Entrepreneurship and Innovation: Public Policy Frameworks." *Journal of Technology Transfer* 37(1): 1–17.

Berger, Suzanne. 2013. *Making in America.* Cambridge, MA: MIT Press.

Berger, Suzanne, and MIT Industrial Performance Center. 2005. *How We Compete: What Companies around the World Are Doing to Make It in Today's Global Economy.* New York: Currency Doubleday.

Boroush, Mark. 2010. "NSF Releases New Statistics on Business Innovation." *InfoBrief* NSF 11-300, National Science Foundation, Directorate for Social, Behavioral, and Economic Sciences. October.

Bureau of Labor Statistics. 2012. "Manufacturing Compensation Costs in Foreign Countries and in the United States, 2011." December 21. http://www.bls.gov/opub/ted/2012/ted_20121221.htm.

Bureau of Labor Statistics. n.d "Manufacturing Sector and NIPA-level Manufacturing Industries KLEMS Multifactor Productivity Tables by Industry." Available at http://www.bls.gov/mfp/mprdload.htm.

Chesbrough, Henry. 2004. "Managing Open Innovation." *Research-Technology Management* 47(1): 23–6.

Christensen, Clayton M., Christopher S. Musso, and Scott D. Anthony. 2004. "Maximizing the Returns from Research." *Research–Technology Management* 47(4): 12–18.

David, Paul A., and Bronwyn H. Hall. 2000. "Heart of Darkness: Modeling Public-Private Funding Interactions inside the Black Box." NBER Working Paper 7538.

Fuchs, Erica R. H., Frank R. Field, Richard Roth, and Randolf Kirchain. 2011. "Plastic Cars in China? The Significance of Production Location over Markets for Technology Competitiveness in the United States versus the People's Republic of China." *International Journal of Production Economics* 132(1): 79–92.

Gallaher, Michael, Jeffrey Petrusa, Alan O'Conner, and Stephanie Houghton. 2007a. *Economic Assessment of the Technical Infrastructure Needs of the Biopharmaceutical Industry.* Planning Report 07-1, prepared by RTI for National Institute of Standards & Technology Gaithersburg, MD: National Institute of Standards and Technology. http://www.nist.gov/director/planning/upload/report07-1.pdf.

Gallaher, Michael P., Brent R. Rowe, Alex V. Rogozhin, Stephanie A. Houghton, J. Lynn Davis, Michael K. Lamvik, and John S. Geikler. 2007b. *Economic Impact of Measurement in the Semiconductor Industry.* Prepared for RTI International for National Institute of Standards & Technology. Planning Report 07-2. Gaithersburg, MD: National Institute of Standards and Technology. http://www.nist.gov/director/planning/upload/report07-2.pdf.

Goffin, Keith, and Ursala Koners. 2011. "Tacit Knowledge, Lessons Learnt, and New Product Development." *Journal of Product Innovation and Management* 28(2): 300–318.

Griliches, Zvi. 1986. "Productivity, R&D, and Basic Research at the Firm Level in the 1970s." *American Economic Review* 76(1): 141–54.

Hecker, Daniel E. 2005. "High-Technology Employment: A NAICS-based Update." *Monthly Labor Review* 128(7): 57–72.

IPC. 2012. "World PCB Production Report for the Year 2012." (Summary global production share chart in http://www.youtube.com/watch?v=VvB4UJf0nZM).

Jones, Charles I., and John C. Williams. 1998. "Measuring the Social Returns to R&D." *Quarterly Journal of Economics* 113(4): 1119–35.

Jones, Charles I., and John C. Williams. 2000. "Too Much of a Good Thing? The Economics of Investment in R&D." *Journal of Economic Growth* 5(1): 65–85.

Link, Albert N., and John T. Scott. 2011. *The Theory and Practice of Public-Sector R&D Economic Impact Analysis.* Planning Report 11-1. Gaithersburg, MD: National Institute of Standards and Technology. http://www.nist.gov/director/planning/upload/report11-1.pdf.

Krugman, Paul. 2013. "Sympathy for the Luddites." *New York Times*, June 13.

Kurfess, Thomas R. 2011. "The Growing Role of Electronics in Automobiles: A Timeline of Electronics in Cars." June 2. http://www.chicagofed.org/digital_assets/others/events/2011/automotive_outlook_symposium/kurfess_060211.pdf.

Lucas, Robert E. 2009. "Trade and the Diffusion of the Industrial Revolution." *American Economic Journal: Macroeconomics* 1(1): 1–25.

Mansfield, Edwin. 1980. "Basic Research and Productivity Increase in Manufacturing." *American Economic Review* 70(5): 863–73.

National Institute of Standards and Technology. (n.d) NIST Strategic Planning Studies. Available at http://www.nist.gov/director/planning/planning.cfm.

Macher, Jeffrey T., David C. Mowery, and Alberto Di Minin. 2008. "Semiconductors." Chap. 3 in *Innovation in Global Industries*, edited by Jeffrey T. Macher and David C. Mowery. Washington, DC: National Academies Press.

McKinsey Global Institute. 2012. *Manufacturing the Future: The Next Era of Global Growth and Innovation.* http://www.mckinsey.com/insights/manufacturing/the_future_of_manufacturing.

Muro, Mark, and Scott Andes. 2013. "Foxconn Sends a Manufacturing Message with New Pennsylvania Plant." Brookings Institution, November 25, 2013 http://www.brookings.edu/blogs/the-avenue/posts/2013/11/25-foxconn-pennsylvania-muro-andes.

National Science Foundation. n.d. "National Patterns of R&D Resources: 2010–11 Data Update." http://www.nsf.gov/statistics/nsf13318/content.cfm?pub_id=4268&id=2.

Petrick, Irene. 2009. "Supply Chain Globalization: How Surviving SMEs Can Position Themselves for the Future." Chap. 8 in *Manufacturing a Better Future for America*, edited by Richard McCormick. Washington, DC: Alliance for American Manufacturing.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5, Part 2): S71–S102.

Schumpeter, Joseph A. 1950. *Capitalism, Socialism, and Democracy,* 3rd ed. New York: Harper and Row.

Solow, Robert. 1956. "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics* 70(1): 65–94.

Solow, Robert M. 1957. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics* 39(3): 312–20.

Stokes, Donald E. 1997. *Pasteur's Quadrant: Basic Science and Technological Innovation.* Washington, DC: Brookings Institution Press.

Tang, Christopher S., and Joshua Zimmerman D. 2009. "Managing New Product Development and Supply Chain Risks: The Boeing 787 Case." *Supply Chain Forum* 10(2): 74–86.

Tassey, Gregory. 1992. *Technology Infrastructure and Competitive Position.* Norwell, MA: Kluwer.

Tassey, Gregory. 2005. "The Disaggregated Technology Production Function: A New Model of University and Corporate Research." *Research Policy* 34(3): 287–303.

Tassey, Gregory. 2007. *The Technology Imperative.* Northampton, MA: Edward Elgar.

Tassey, Gregory. 2008. "Modeling and Measuring the Economic Roles of Technology Infrastructure." *Economics of Innovation and New Technology* 17(7–8): 615–29.

Tassey, Gregory. 2010. "Rationales and Mechanisms for Revitalizing US Manufacturing R&D Strategies." *Journal of Technology Transfer* 35(3): 283–333.

Tassey, Gregory. 2013a. "Beyond the Business Cycle: The Need for a Technology-Based Growth Strategy." *Science and Public Policy* 40(3): 293–315.

Tassey, Gregory. 2013b. "Technology Life Cycles." In *The Encyclopedia of Creativity, Invention, Innovation, and Entrepreneurship,* edited by Elias Carayannis and David Campbell. New York, NY: Springer Science + Business Media.

US Census Bureau. n.d. "Trade in Goods with Advance Technology Products." http://www.census.gov/foreign-trade/balance/c0007.html.

Welser, Jeff. 2008. "NIST & SRC Nanoelectronics Research Initiative: Partnership for Innovation." October. http://www.nist.gov/director/vcat/upload/NIST_SRC_NRI.pdf.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 49–72

Management Practices, Relational Contracts, and the Decline of General Motors

Susan Helper and Rebecca Henderson

General Motors was once regarded as the best-managed and most successful firm in the world. However, between 1980 and 2009 GM's US market share fell from 46 to 20 percent, and in 2009 the firm went bankrupt. (Figure 1 shows the changing market shares of GM and its main competitors over time.) We argue that the conventional explanation for this decline—namely high legacy labor and healthcare costs—is seriously incomplete, and that GM's share collapsed for many of the same reasons that many highly successful American firms of the 1960s were forced from the market, including a failure to understand the nature of the competition they faced and an inability to respond effectively once they did. We focus particularly on the problems GM encountered in developing the relational contracts essential to modern design and manufacturing. We discuss a number of possible causes for these difficulties: particularly, GM's historical practice of treating both its suppliers and its blue collar workers as homogeneous, interchangeable entities; its view that expertise could be partitioned with minimal overlap of knowledge amongst functions or levels in the organizational hierarchy; and its faith that decisions should be based largely on well-defined financial criteria. We suggest that GM's experience may have important implications for our understanding of the role of management in the modern, knowledge-based firm and for the potential revival of manufacturing in the United States.

■ *Susan Helper is AT&T Professor of Economics, Case Western Reserve University, Cleveland, Ohio. She is currently serving as Chief Economist of the US Department of Commerce. Rebecca Henderson is the John and Natty McArthur University Professor, Harvard University, Cambridge, Massachusetts. Their emails are susan.helper@case.edu and Rhenderson@hbs.edu. The views expressed are those of the authors alone.*

http://dx.doi.org/10.1257/jep.28.1.49 doi=10.1257/jep.28.1.49

Figure 1
Market Shares in the US Automotive Industry over Time

Source: Wards Automotive, http://wardsauto.com/public-data.

What led to the decline of General Motors? The answer is worth exploring, not only because General Motors was arguably the most important firm of its time, but also because over the same period so many other once-great American firms also saw their positions erode dramatically. The US steel, electronics, and apparel industries came very close to collapse. Only 13 percent of the Fortune 500 in 1963 were still amongst the 500 a half-century later. If we are to develop an understanding of how US manufacturing might be revived, it is important to understand what led to its decline.

In this paper, we investigate this issue through an exploration of the decline of General Motors. We begin by laying out a number of alternative explanations. We suggest that while GM was indeed handicapped by high legacy costs, its decline was driven largely by the poor design and inferior quality of its cars and the low productivity of its operations. The principal puzzle of GM's failure is why the firm was seemingly unable to adopt the managerial practices that enabled its Japanese competitors—particularly Toyota—to introduce cars of much higher quality and much better design at significantly lower cost, even though GM was, at least initially, much richer than its rivals. The puzzle is deepened by the fact that the nature of the practices underlying Japanese success was extensively documented in both the scholarly and business press, and GM had firsthand experience with these practices through a joint venture with Toyota at a production facility in California starting in the middle of the 1980s.[1]

We identify two principal answers to this puzzle. First, we argue that the historical success of General Motors led the firm's senior managers to deny and/or misperceive the nature of the threat presented by Japanese competition for much of the 1970s and 1980s. The second answer focuses on the difficulties that GM experienced in the 1990s once the firm had made the decision to adopt Toyota's managerial practices: we suggest that it took time for GM to understand exactly what Toyota was doing; and then problems in building new relational contracts greatly slowed GM's efforts to respond effectively, either through innovation or by imitating Toyota's efforts.[2] We close with a discussion of the implications of this history for efforts to revive American manufacturing.

Parsing Alternative Explanations for the Decline of General Motors

Perhaps the most popular explanation for the failure of General Motors is that decades of overly generous union contracts put it at an overwhelming cost disadvantage (Ingrassia 2010). GM did have higher labor costs than Toyota. For example, the Associated Press (2007) reported that because of "legacy" health care and pensions owed to retired workers, labor costs at General Motors were as high as $73/hour while GM's Japanese competitors' costs were roughly $48/hour. Another calculation suggests that legacy costs at General Motors were about $1,600 per car in 2005 (Welch and Beucke 2005). However there are several reasons for believing that this cost disadvantage was not the only issue that caused GM's difficulties.

First, GM's legacy costs were high largely *because* of GM's declining market share. If GM had maintained its 1980 US market share until 2009, for example, its per-car labor costs would have fallen by one-half.

Second, poor quality and poor design were at least as significant a problem as that of legacy costs. Throughout the 1980s and 1990s, consumers complained that American cars suffered from noise, vibration, and harshness and from poor ride quality. When Toyota and General Motors were running a joint venture together

[1] GM also faced a challenge from European automakers during this period (1980–2009). However, the Europeans' US market share was well under 25 percent of that held by the Japanese, and we do not address it here.

[2] It seems unlikely that Toyota's approach to the market was the "one best way" to compete; it is entirely possible that GM could have responded to Toyota's challenge by developing its own innovative methods of designing and producing cars. In what follows, we focus on GM's difficulties in imitating Toyota because a) GM managers starting in the late 1990s often used Toyota as a benchmark and b) Toyota's practices provide a clear alternative; had GM been able to adopt them easily, its performance would have improved significantly compared to the decades of stagnation we in fact saw. For more discussion of the issue of alternatives to "Toyotism," see Freyssenet, Mair, Shimizu, and Volpato (1998).

Table 1
The Productivity of GM's Framingham Assembly Plant versus the Toyota Takaoka Assembly Plant, 1986

	GM Framingham	*Toyota Takaoka*
Gross assembly hours per car	40.7	18.0
Adjusted assembly hours per car	31	16
Assembly defects per 100 cars	130	45
Assembly space per car (square feet/year)	8.1	4.8
Inventories of parts (average)	2 weeks	2 hours

Source: Womack, Jones, and Roos (1990).
Notes: "Gross assembly hours per car" measures the total time required to assemble the car model made at each plant. "Adjusted hours" is the time required to do a standard set of assembly activities. "Assembly defects" refers to quality problems related to the assembly process (as opposed to those caused by defective supplier parts, for example). "Inventories" measures how long it would take for the plant to exhaust its parts supply at its current production rate.

in the late 1980s and early 1990s, those cars coming off the line with the Toyota nameplate commanded more than a 20 percent premium in the marketplace over their nearly identical GM brethren (Sullivan 1998). More generally, in 2000 all GM cars sold on average for $3,000 less than Toyotas or Hondas of comparable size and equipment (Train and Winston 2007), implying that GM's pricing disadvantage was greater than its legacy cost disadvantage. It was not until 2012 that Chevrolet's quality rankings began to approach those of Toyota and Honda.

Third, General Motors appears to have been significantly less productive than its rivals in nearly every aspect of its operations. For example, Clark and Fujimoto (1991) found that in the late 1980s the Japanese took 1.7 million adjusted engineering hours to develop a $14,000 car, while their US competitors took an average of 3.2 million hours. GM's assembly operations were also persistently less productive than those of its rivals. For example, Table 1 shows a comparison developed by the MIT International Motor Vehicle Program between the GM plant in Framingham, Massachusetts, and the Toyota Takaoka plant in Japan. The first row shows gross hours required to assemble a car, while the second row shows that, even after adjustments to ensure "apples to apples" comparisons, it was taking GM nearly double the number of hours to assemble a car (Womack, Jones, and Roos 1990).

Last but not least, direct labor costs (including legacy costs) were only 10 percent of total costs at General Motors, while purchased parts were 70 percent (Helper and MacDuffie 2008), and there is a great deal of evidence that GM's management of its supplier network was significantly less effective than that of its Japanese rivals. For example, Clark and Fujimoto (1991) estimated that supplier contributions accounted for one-third of the difference in the Japanese automakers' advantage over their US counterparts in total engineering hours required to develop a new car, while defect rates of parts supplied by Japanese companies were on the order of one-tenth the rate of those supplied by US firms (Cusumano and Takeishi

1991). Similarly, in a sample of US automotive suppliers selling to both Toyota and US automakers, Dyer and Hatch (2006) found that from 1990 to 1996 the average supplier reduced defects by 50 percent for Toyota, but only 26 percent for their largest US customer.

These failures are all the more striking in that they were widely documented as they were occurring. Industry observers deplored General Motors' failure to introduce attractive small cars throughout the 1970s and 1980s.

In addition, from 1984 on, General Motors had privileged access to Toyota's techniques through NUMMI (New United Motor Manufacturing, Inc.), a joint venture that GM formed with Toyota precisely to gain insight into Toyota's capabilities. NUMMI was located in GM's Fremont (California) plant, which had been closed two years before. The plant had been one of the least productive in GM, and had been famous for its troubled labor relations. Unexcused absenteeism often exceeded 20 percent, the plant had accumulated a backlog of over 1,000 grievances, and the union–management relationship was described by a top union member as an "ongoing war" (Brown and Reich 1989). Despite this background, Toyota agreed to rehire the entire union hierarchy, and when production began, 99 percent of the assembly workers and 75 percent of the skilled trades workers were former GM-Fremont employees and UAW members (Adler, Kochan, MacDuffie, Pil, and Rubenstein 1997; Brown and Reich 1989). NUMMI went on to reach levels of productivity and quality comparable to those of Toyota's Japanese plants (Womack, Jones, and Roos 1990). Similarly, in 1983, GM invested more than $5 billion in Saturn, a homegrown attempt to reinvent GM's product design procedures, manufacturing techniques, and labor relations. The attempt was successful from many perspectives, and yet it, too, seems to have had only a minimal impact on the functioning of GM as a whole (Rubenstein and Kochan 2001).

Moreover, during the key period of the 1980s, General Motors was not short of cash. Between 1980 and 1985, the firm spent over $45 billion on acquisitions and automation (Keller 1989), a sum that at the time was more than the combined market value of Toyota and Honda. Nor does it seem plausible that the firm's disadvantage was a function of unique features of Japanese culture or governance, or the uniquely recalcitrant nature of GM's union and workforce, since the successes at both NUMMI and Saturn makes these explanations implausible.

Another popular explanation for the troubles at the firm has been that the firm was simply badly managed (Keller 1989; Taylor 2010), but contemporaneous assessments suggest that the senior managers at General Motors were at least as capable as those at other firms. When Roger Smith was chief executive officer of GM in the 1980s, he was named *Automotive Industries* Man of the Year, *Advertising Age's* Ad Man of the Year, and *Financial World* magazine's Chief Executive Officer of the Year, and he was also designated by the Gallagher Report as one of the 10 best executives in the United States (Finkelstein 2003). Yet at the same time, Smith failed to invest in the development of small cars, spent hundreds of millions of dollars on "high technology" acquisitions that did little or nothing for GM, and (apparently) did little to diffuse the lessons of NUMMI across the company.

What can explain these patterns? Why were the product development processes at General Motors so slow and expensive and its design capabilities so inferior? Why were GM's supplier network and assembly operations so much less productive than those of its rivals, and why was the quality of their output so much lower? Why did these trends persist for so long?

One stream of work argues that these problems are failures of *perception* and *motivation*. Problems of perception—or of the failure to recognize that the world is changing—flow from the fact that senior managers tend to become overly reliant on the mental models and beliefs that undergirded the firm's success in the first place. Problems of motivation—or of an unwillingness to act even once the need to change has been recognized—can arise when the selection environment[3] in an industry is weak and/or senior managers are insufficiently motivated to act in the interests of the firm.

In the case of General Motors, it is hard to tease these two issues apart without more detailed data. But because the American automobile industry in the 1950s and 1960s was a reasonably collusive oligopoly, it seems plausible that when Toyota and the other Japanese firms first began to make serious inroads into the US market, GM's senior management had little experience of intense competition. Perhaps as a result, throughout the 1970s they blamed the success of their Japanese rivals on Japan's low labor costs and inferior working conditions, and insisted that their own small cars were well designed and competitively viable, despite widespread ridicule in the industry press (Behr 1981). Qualitative accounts stress the ways in which the firm's past success (and the industry's insularity, due to its concentration in Detroit) led managers to dismiss Japanese inroads as reflecting the odd—even aberrant—preferences of consumers on the East and West coasts, and to claim that data demonstrating Toyota's superior productivity and quality was incorrect or misleading (Keller 1989; Taylor 2010; Yates 1983). A historically weak selection environment may thus have handicapped the firm when it had to respond to much stronger competition.

Even after managers at General Motors became convinced that Toyota was indeed doing "something different" in its factories—a development many industry observers credit to the appointment of Roger Smith as GM's chief executive officer in 1981—they appeared to have believed that the essence of Toyota's advantage lay in tools like the fixtures designed to change stamping dies rapidly, or in the use of "just in time" inventory systems, rather than in the management practices that made it possible to develop and deploy these techniques. For example, Jeffrey Liker, a professor and consultant to GM since the 1980s, reported (in "This American Life," 2010):

> One of the GM managers was ordered, from a very senior level—(it) came from a vice president—to make a GM plant look like NUMMI. And he said, "I want

[3] "Selection environment" refers to the socioeconomic environment created by outside actors (such as suppliers and competitors) and external social and legal institutions (such as regulators and industry standards and norms). In weak selection environments, managers may not have had experience in dealing with stiff competition, quality problems, or more stringent environmental regulations. For an overview of this work, see Gibbons and Henderson (2013) and the references there.

> you to go there with cameras and take a picture of every square inch. And whatever you take a picture of; I want it to look like that in our plant. There should be no excuse for why we're different than NUMMI, why our quality is lower, why our productivity isn't as high, because you're going to copy everything you see. . . . Immediately, this guy knew that was crazy. We can't copy employee motivation; we can't copy good relationships between the union and management. That's not something you can copy, and you can't even take a photograph of it.

Outside General Motors, the idea that one could copy NUMMI—or indeed any of Toyota's advantages—by simply copying the physical plant was seen to be mistaken quite early. "The Machine that Changed the World" (Womack, Jones, and Roos 1990)—a report from the MIT automobile project that incorporated detailed productivity data from across the world and in which GM had actively participated, providing detailed data on GM plants and attending meetings from 1985 on—was published in 1990, while Clark and Fujimoto's (1991) detailed study of product development across the world automobile industry came out the following year. These studies (along with many others) documented the very different ways in which Toyota managed product design, assembly, and its supplier network.

Despite this flood of research, it took General Motors more than two decades to imitate Toyota's practices consistently. While problems in perception and motivation are certainly plausible explanations of why GM took so long to internalize fully the idea that Toyota was indeed doing something differently, they are much less satisfactory explanations for why it took so long for the firm to respond effectively once it did have this realization. Here we make the case that GM struggled for so long because Toyota's practices were rooted in the widespread deployment of effective *relational contracts*—agreements based on subjective measures of performance that could neither be fully specified beforehand nor verified after the fact and were thus enforced by the shadow of the future—and that GM's history, organizational structure, and managerial practices made it very difficult to maintain these kinds of agreements either within the firm or between the firm and its suppliers.

We begin by describing the very different management practices General Motors and Toyota employed to design and manufacture automobiles in the 1970s and early 1980s, paying particular attention to the ways in which GM's were predicated on a view of workers, suppliers, and even white collar employees as commodities whose work could be fully controlled by experts through the use of careful specifications and the spot market, while Toyota's practices were critically dependent on joint problem solving across boundaries of all kinds, and thus on the existence of strong relational contracts.

Comparing Managerial Practices at Toyota and General Motors

Automotive Assembly

In the 1960s and 1970s, jobs on the General Motors assembly line were very narrowly defined; a worker would perform the same set of tasks—for example, screwing in several bolts—every 60 seconds for eight to ten hours per day. Workers were not expected or encouraged to do anything beyond this single task. Responsibility for the design and improvement of the assembly system was vested firmly in the hands of supervisors and manufacturing engineers, while vehicle quality was the responsibility of the quality department, which inspected vehicles as they came off the assembly line. GM's managers were notorious for believing that blue collar workers had little—if anything—to contribute to the improvement of the production process (Womack, Jones, and Roos 1990; Adler and Borys 1992). Relationships between blue collar workers and local management were actively hostile. For example Joel Smith, a NUMMI worker interviewed by Adler (1993) described life "in the old days":

> In the old days, we fought for job security in various ways: "Slow down, don't work so fast." "Don't show that guy next door how to do your job—management will get one of you to do both of your jobs." "Every now and then, throw a monkey wrench into the whole thing so the equipment breaks down—the repair people will have to come in and we'll be able to sit around and drink coffee. They may even have to hire another guy and that'll put me further up on the seniority list.
>
> Management would respond in kind: "Kick ass and take names. The dumb bastards don't know what they're doing." . . . Management was looking for employees who they could bully into doing the job the way they wanted it done. The message was simply: "If you don't do it my way I'll fire you and put somebody in who will. There are ten more guys at the door looking for your job."

Jobs on Toyota's production line were even more precisely specified: for example, standardized work instructions specified which hand should be used to pick up each bolt. However, Toyota's employees had a much broader range of responsibilities.[4] Each worker was extensively cross-trained and was expected to be able to handle six to eight different jobs on the line. They were also responsible for both the quality of the vehicle and for the continual improvement of the production process itself. Each worker was expected to identify quality problems as they occurred, to pull the "andon" cord that was located at each assembly station to

[4] The workforce management techniques employed by Toyota have been extensively studied by labor economists and specialists in industrial relations. Together they are often called "high performance work systems." There is no single definition of "high-performance work system," but three overarching elements have been identified in the literature. In general, firms with high-performance work systems 1) implement effective incentive systems, 2) pay a great deal of attention to skills development, and 3) use teams and create widespread opportunities for distributed communication and problem solving. For examples, see Kochan, Katz, and McKersie (1986); MacDuffie (1995); Huselid (1995); Ichniowski, Shaw, and Prennushi (1997); Pfeffer (1998); Appelbaum, Bailey, Berg, and Kalleberg (2000); and Black and Lynch (2001). Despite Toyota's general adherence to the principles above, many workers on Toyota's lines were temporary employees with little job security, and in some cases the pace of work was punishingly fast (see for example, Parker and Slaughter 1994).

summon help to solve them in real time, and if necessary to pull the andon cord again to stop the entire production line. Workers were also expected to play an active role in teams that had responsibility for "continuous improvement" or for identifying improvements to the process that might increase the speed or efficiency of the line. As part of this process, workers were trained in statistical process control and in experimental design.

General Motors faced two problems in imitating these practices. The first was in understanding exactly how they were constructed. Milgrom and Roberts (1990, 1995) argue, for example, that diffusion of these techniques was hampered by the fact that they display strong complementarities, and in two papers, Bloom and his collaborators (Bloom, Eifert, McKenzie, Mahajan, and Roberts 2013; Bloom, Sadun, and Van Reenen 2013) make the case for thinking about these practices as "management technologies" that face similar barriers to diffusion as do more conventional technologies.

The second problem was that these practices were necessarily rooted in relational contracts. It was, for example, very difficult to specify the exact circumstances under which a worker should pull the andon cord, or what behaviors constituted being an effective team member. Shutting down the line for a popular model could cost $10,000 in lost profits per minute (Helper 2011), so management setting up this system needed to be confident that a worker deciding to pull the andon cord would have both the knowledge and the incentive to exercise sophisticated judgment. Conversely, workers would only pull the cord if they were confident that an appropriate relational contract was in place (Gibbons and Henderson 2013). Similarly MacDuffie's (1997) detailed description of the practices underlying shop-floor problem solving in the industry suggests that successful process quality improvement depended on processes that allowed for the inclusion of multiple perspectives on any single problem, the use of problem categories that were "fuzzy," and the development of a common language for discussing problems. It seems implausible that employees could be motivated to participate in these kinds of activities through the use of formal contracts that specified in advance every kind of quality problem and its appropriate response, especially given the many potential problems that could arise in an auto assembly plant, which typically has over 500 different work stations each performing a different set of tasks.

Supply Chain Management

US automakers' relationships with outside suppliers from the 1950s to the early 1990s were also very different from those of their Japanese competitors. They were characterized by short-term—usually one-year—contracts, arm's-length relationships, and a reliance on as many as six to eight suppliers per part. Purchase decisions were driven overwhelmingly by price and were governed by written contracts that involved precise specification of the purchased part. There was little communication between suppliers and either the central engineering groups who designed the parts or the assembly plants responsible for using them—a reflection of a deeply held belief at General Motors that experts should do the planning and designing while implementation should be accomplished through a detailed division of labor (MacDuffie and Helper 1997). While this belief did not promote quality, it did facilitate the maintenance of "spot" relationships with suppliers.

In the Japanese automobile industry, in contrast, firms were much more likely to enter into long-term relationships with suppliers and were much less likely to switch suppliers because of small differences in quoted price. Because of their belief in *genchi gembutsu* (that detailed knowledge of context is valuable), Toyota and its competitors typically did not develop a detailed design for many parts in a new model. Instead, they would specify the part's exterior dimensions and its performance characteristics, and allow a specialist supplier to design the part instead to best match the supplier's production process (Cusumano 1985; Smitka 1991; Nishiguchi 1994). Toyota's use of "knowledge overlap" (Takeishi 2002) between its engineers and its suppliers' engineers not only allowed for better problem-solving, but also helped Toyota ensure that its suppliers remained near the production frontier. In contrast, a top purchasing manager at General Motors explained in 1993: "GM doesn't need to understand the technologies that our suppliers use—we let the market tell us" (Susan Helper, unpublished interview).

Suppliers to the Japanese car firms were significantly more likely to invest in organizational capabilities such as quality training and maintaining a product-design staff, allowing them to engage in activities such as "value analysis" and "value engineering," which are techniques that involve examining the contribution to cost and functionality of each aspect of a component's design. The Japanese automobile firms and their suppliers also invested in organizational mechanisms designed to increase information flow, which meant that the parties came to deeply understand each other's products and processes. These practices promoted continuous improvement, allowing Japanese suppliers to reduce prices every year for decades while remaining profitable, and enabled the Japanese automotive companies to be confident in producing "just-in-time" rather than holding costly inventory (Lieberman and Demeester 1999; Lieberman and Asaba 1997; Lieberman, Helper, and Demeester 1999).

US and Japanese automobile firms also developed very different ways of handling the design changes that were frequently made as responses to unforeseen interaction problems. Soderberg (1989), for example, estimates that on average, each part was changed at least once in the industry during the 1980s. Thus suppliers were often not making precisely the part they were originally contracted to make. In US practice, changing specifications meant legally changing the contract, and suppliers were often able to extract high prices for making these changes. In contrast, Japanese manufacturers simply asked suppliers to make the change; both parties trusted each other to "sort things out" later. As one supplier to several automakers said, "Honda cares about making the part fit the car, while Ford cares about making the part fit the blueprint" (MacDuffie and Helper 1997).

In consequence, the relationships between Japanese firms and their suppliers were deeply rooted in relational contracts. Suppliers "know that as long as they make a good-faith effort to perform as they should, the assembler will ensure that they receive a reasonable return on their investment" (Womack, Jones, and Roos 1990), and as long as the supplier continued to meet the automaker's expectations, the supplier could count on the relationship continuing indefinitely. Smitka (1991) describes these arrangements as "governance by trust." This reliance did not mean that the Japanese paid less attention to performance management than the American firms. Toyota's relational contract with suppliers was not "a cozy relationship," as one manager of a supplier company pointed out (Helper and Sako 1995). The firm pushed its suppliers very hard to reduce costs and avoid defects; it reduced the market share of suppliers who did not meet these strict goals and exited the relationship completely if improvement was not forthcoming. In fact, Honda and Toyota collected more data about supplier performance than GM did during this period. However, in contrast to a "make the numbers" culture (Repenning and Henderson 2010), Toyota did not award high-powered incentives to firms or individuals that performed well on one or more of these targets. Instead, the objective data was used as an aid to understanding the "root causes" of problems, in conjunction with subjective data and intuition (MacDuffie and Helper 1997).

Product Design and Development

Before 1984, product development within General Motors was managed by three separate organizations: a car division, such as Buick, that was responsible for the car's design; Fisher Body, which was responsible for the detailed engineering; and GMAD (the General Motors Assembly Division), which would modify plants and equipment to prepare for the new model and ultimately assemble the car. There was only very limited interaction between the three. All divisions reported ultimately to the president, who was responsible for arbitrating disputes. According to Keller (1989, pp. 100, 101, 106), "Each of the three (organizations) viewed itself as a separate entity with the necessity of protecting its own autonomy." Consultants hired to evaluate the process found that "the bureaucracy was a virtual quicksand bog of procedures" in which "individuals were not held accountable for the decisions they made."

Then in 1984, General Motors was reorganized into two divisions: "BOC", which was composed of the Buick, Oldsmobile, and Cadillac car divisions, and "CPC", which was composed of Chevrolet, Pontiac, and GM Canada. Fisher Body and GMAD were broken up and combined with BOC and CPC. The apparent intention was to streamline and integrate new product development, but the reorganization created considerable confusion and did not noticeably improve performance. The informal agreements—or in our terms, the relational contracts—that lower-level GM managers had established with each other were purposely broken up, either because they were not valued or because they were assumed to be an active impediment to improvement. However GM's formal organization was so cumbersome that work proceeded even more slowly, as managers did not know if they could trust their counterparts to ignore some of the red tape (Keller 1989), an observation consistent with Kaplan and Henderson's (2005) suggestion that the need to remake relational contracts may be a significant barrier to the ability to develop new ones.

Both divisions relied on "light-weight project teams"—coordinating mechanisms in which the project manager attempted to coordinate the work of the multiple functions whose work was critical to product design but without the benefit of any real authority over the team's members (Clark and Fujimoto 1991). Within this structure, key decisions about product design appear to have been driven as much by the finance function as by the project leader, and engineers and process designers appear to have focused as much on the health of their own local organizations as on the strength of the design process itself.

For example, in the mid-1990s, General Motors previewed a widely-heralded concept vehicle called the Aztek. As the *Washington Post* pointed out: "The concept car actually did something few GM designs do: arrive before a trend—this time, the crossover SUV that combines the attributes of a truck and a passenger car" (Weissman 2005). However, the production car "represented all that is wrong with GM's design process," according to a GM executive. "The penny-pinchers demanded that costs be kept low by putting the concept car on an existing minivan platform. That destroyed the original proportions and produced the vehicle's bizarre, pushed-up back end." According to *Motoramic* magazine, "Tight budgets and boardroom dominance of manufacturing over design meant the underlying bits of the Aztek were set in stone before stylists ever lifted a pencil; a wheezy 3.4-liter V-6, a frame that was about 15 percent too big, and no freedom for designers to alter any major components killed whatever visual appeal and sporty pretentions the original shape held" (Hyde 2013). The car won several awards for ugliness, never reached its modest sales targets, and was soon taken out of production.

In contrast, product design and development at Toyota was managed through tightly knit, dedicated "heavyweight" project development teams. Team leaders were managers of long experience who had full authority over a team composed not only of representatives from engineering and design, but also from manufacturing, sales, and marketing. They had responsibility for the entire lifecycle of the product: from concept through detailed engineering to manufacturing and commercial launch. For example, one team defined its goal as designing a car that felt like "a rugby player in a business suit," a concept that informed every aspect of the subsequent process (Clark and Fujimoto 1991).

While performance at General Motors, when it was judged at all, seems to have been judged on the basis of well-defined rules or easily observable metrics such as whether individuals met pre-specified deadlines, performance at Toyota was judged on the basis of the performance of the team as a whole.[5] More broadly, the Toyota Production System was embedded in the Toyota Management System via a set of goals that linked shop floor activities and targets to the objectives of both the senior

[5] Wright (1979) reports one extreme example of rule following. He reports that at General Motors in the 1970s it was considered a great honor for a junior executive to be chosen to run the slide presentation at board meetings but that the executive's career could be ended if he put a slide in the projector incorrectly.

team and functional managers. These goals were jointly determined through lively communication across multiple levels of the organization, again suggesting that the white collar workforce was shot through with relational contracts (Liker 2004; Sako 2006).

Why Did General Motors Struggle to Adopt Toyota's Management Practices?

The General Motors share of the US market began to plummet around 1980, but it took the firm more than 20 years to adopt Toyota's techniques. For example, MacDuffie (1995) and MacDuffie, Sethuranman, and Fisher (1996) surveyed auto plants around the world, and they found only slow progress toward using a Toyota-style production system from 1986 to 1993. Similarly, GM suppliers responding to Susan Helper's surveys of auto suppliers reported that in 1984 they did not expect their relationship to last into the future, and that they did not expect to provide or receive "help not required by the contract." By 1989, GM had the highest scores among the Detroit three on items such as expecting that their customer would help them rather than switch immediately to a rival offering a better deal; but by 1993, suppliers (again) felt that GM could not be trusted (Helper 1994; Helper and Sako 1995). Why was this?

First, it took General Motors some time to understand exactly what Toyota was doing and to attempt to implement the full bundle of practices necessary to successfully imitate its Japanese rivals. For example, managers at GM's Van Nuys plant introduced elements of the practices that had been implemented at NUMMI, but they resulted in significantly less improvement, and Van Nuys was soon closed ("This American Life" 2010).

Second, General Motors had great difficulty building the relational contracts on which these practices relied. The research literature has typically assumed that if a relational contract is mutually beneficial, it will naturally emerge (Levin 2003), but a body of work has begun to explore the conditions under which relational contracts may be hard to build. We draw on Gibbons and Henderson's (2013) summary of this research and explore the degree to which it yields insight into GM's decline.

Gibbons and Henderson (2013) outline three broad classes of explanation for why relational contracts may be difficult to build. First, there may be *unobserved heterogeneity* in the costs of using relational contracts. For example, if the contracting parties have heterogeneous discount rates, then an equilibrium may emerge in which more patient players form relational contracts while less patient players do not. For example, in a model of supply relationships, Board (2011) finds that the optimal relational contract has an insider-outsider form where some suppliers are "insiders" who are used routinely and the rest are "outsiders" who are never used, even in periods where they are more efficient. He shows that if the parties become more patient, the set of insiders grows and efficiency improves. Similarly, if the principal has private information about his or her temptations to renege, then incentives for a relational contract will only strengthen gradually, as the agent becomes convinced that the principal is unlikely to defect (Halac 2012). Second, problems of *credibility* may arise when it is hard to tell whether a party is taking a hard line in a relational situation because it wishes to take advantage of the other party or because a poor state of the world is genuinely forcing them to take a hard line (Chassang 2010; Li and Matouschek 2013). Third, problems of *clarity*, or the simple inability of principal and agent to communicate with each other, may also create problems.

Exploring the history of General Motors suggests all three of these explanations may have played a role in shaping the firm's history. Up until the 1990s, GM was receiving oligopoly rents. In this situation, arm's-length relations maximized profits compared to relational contracts because they made suppliers easily replaceable, thus reducing suppliers' ability to bargain for a share of these rents (Helper and Levine 1992). Thus, historically it may have been the case that GM's returns to adopting relational contracts were lower than those of its rivals. As these oligopoly rents shrunk, however, GM's long-term payoffs from adversarial behavior fell, and it seems plausible that the returns to adopting relational contracts became quite high.

At this stage, General Motors faced problems of credibility and clarity. The credibility issue arose because it appears to have been hard for GM to alter past patterns of behavior and hard for GM's workers and suppliers to believe that these patterns were indeed changing. Researchers have long suggested that it takes time to build trust, and that when one or both of the parties to a relationship have a history of "bad behavior," building trust-based relationships may be even harder (Bachmann and Zaheer 2006). GM's stance towards both its blue collar employees and its suppliers had been deeply adversarial. In 1984, for example, the company announced that it was interested in modifying the union contract to support the use of "teams" and of "joint problem solving"—but then a leaked internal memo suggested that GM was planning to use the new contract to reduce headcount (Russo 1984). Throughout the 1980s, many workers became convinced that GM was implementing lean production only as part of an attempt to speed up production and to put employees under even greater pressure (Parker and Slaughter 1994). GM thus faced significant problems in building credibility.

General Motors' practice of focusing almost entirely on short-term financial results, along with its multiple levels of control and its large number of employees and suppliers, also made building credibility difficult. Senior management could announce a commitment to long-term relationships and to building trust, but until these announcements were coupled with similar commitments and altered incentives at the local level, neither blue collar employees nor suppliers appear to have believed that the local managers with whom they had to deal would adhere to a relational contract. At the local level, this may have been partly a function of the lack of accountability or follow-through that seems to have characterized GM. For example, according to "This American Life" (2010), "Weller [a GM manager sent to spread the NUMMI gospel] said some managers were responsive. Others weren't—like the one who asked him to leave his factory after Weller made his presentation about the NUMMI system. When asked why the CEO wouldn't fire a plant manager who resisted a system that was producing better cars at lower costs, Weller said: 'It's a big company . . . and it doesn't work that way.'"

Similarly in 1982, Darrah C. Porter, the executive director for purchasing activities at General Motors, told a reporter for *Iron Age*: "We need to throw off the old shackles of adversarial confrontation and work together in an enlightened era of mutual trust and confidence." However, one purchasing agent was quoted in the same article as saying "I find it hard to stop thinking that efficient purchasing means having a lot of vendors fighting over a job" (requoted in Helper 1987). In another example of changes in desired behavior not being accompanied by changed incentives, the *Wall Street Journal* in 1984 reported that GM wanted suppliers to locate within GM's Buick City complex, to promote better communication. However, GM provided no assurance of future business to suppliers who incurred the significant costs of moving, making suppliers reluctant to relocate (Helper 1987).

More broadly, General Motors' history of market dominance appears to have made the firm very risk averse, perhaps because the firm's extraordinary success made risk avoidance a profit-maximizing strategy for many years. One long-ago incident that might have helped to set the stage for such a policy was the failure of GM's attempt to introduce the copper-cooled engine in 1922. The attempt failed largely because the product was forced on the Chevrolet division in violation of GM's policy of divisional autonomy. However, Alfred Sloan (1963, p. 69) drew a broader lesson: he claimed that "it was not necessary to lead in design or run the risk of untried experiment." This incident apparently had a searing staying power; it was cited as late as 1987 by GM engineers as a reason to avoid technological risks (Helper 1991).

We suspect that problems of clarity, or in communicating the terms of the various relational contracts that General Motors was seeking to put in place, were also central to the firm's difficulties. Tables 2, 3, and 4 summarize both the decision problem and the choices for participants in relational contracts a) underlying the use of the andon cord, b) in supplier relationships, and c) in heavyweight product development teams, respectively. As you can see, successful relational contracts in each area are predicated on a complicated set of mutual understandings that would be difficult for an outsider to understand (or enforce). These tables suggest the immense amount of ambiguous information that had to be successfully communicated before a Toyota-style relational contract could be put in place. Not only did everyone concerned have to learn about the actions that constituted "cooperation"—and to come to believe that these actions were in their own best interests—but they also had to learn about each other's defection temptations in a world in which they might not even know their own.

Toyota (and Honda) were able to establish relational contracts through significant investments in "gift exchange" (Akerlof 1984), which in effect meant giving to their suppliers and workers without formal assurance of any returns. For example many workers initially doubted the credibility of NUMMI's no-layoff commitment, but in 1987 and 1988 when NUMMI was running under 65 percent of capacity, Toyota did not lay off shop-floor workers. Instead, it sent the entire workforce to training classes, took back in-house certain previously contracted maintenance

Table 2
Cooperation, Defection, and Punishment in the Use of the Andon Cord

Agent	Action: Cooperate	Action: Defect	Action: Punish
Worker	1. Pull the andon cord when worker sees a problem. 2. Offer suggestions on improvements to the production process (that might make workers' job redundant).	1a. Never pull the andon cord (out of fear of being punished). 1b. Pull the andon cord to stop the line and avoid work when there is no true problem. 2. Keep improvements hidden from co-workers and managers.	1. Sabotage the manufacturing line. 2. Pull andon cord frequently. 3. Engage in absenteeism.
Supervisor	1. Recognize potential problem when andon cord pulled and aid in problem-solving. 2. Implement improvements without necessarily cutting jobs. 3. Accept authority of work teams to make some shop-floor decisions.	1. Punish workers for pulling andon cord (even appropriately). 2. Cut workforce once they discover potential innovations. 3. Interfere in work teams and override their decisions.	1. Penalize workers (financially or socially) for pulling andon cord. 2. Remove the andon cord.

Source: Gibbons and Henderson (2013).

tasks such as painting, and placed surplus workers into teams that designed the production process for the next model car (Adler, Goldoftas, and Levine 1997). This step built a cycle of cooperation in which the union officials started suggesting some ways of cutting costs, and in turn Toyota set up accounts so that the union leaders could order supplies for their team members without having to file requests through management (as detailed in Adler 1993, pp. 68–69).

General Motors seems to have struggled to develop this dynamic. For example, Steven Bera, a GM executive who was first sent to NUMMI and then sent to a number of other GM plants, said (in "This American Life," 2010) that even after GM plants began to install some of the physical features of Japanese auto plants, "there was no change in the culture. Workers and managers continued their old antagonistic ways. In some of the factories where they installed the andon cord, workers got yelled at when they pulled it." Some plant managers continued to believe that blue collar workers were fundamentally lazy and would pull the andon cord any time they wanted a break and that the blue collar workers lacked the capacity to engage in problem solving or continuous improvement. By and large, the blue collar workforce appears to have doubted that the announced reforms would work, albeit for

Table 3
Cooperation, Defection, and Punishment in Supplier Relationships at Toyota

	Action		
Agent	*Cooperate*	*Defect*	*Punish*
Supplier	1. Invest in engineering and process development skills that will enable the firm to translate approximate specifications into a final part. 2. When problems emerge, work rapidly and effectively to fix them.	1. Fail to invest sufficient time or attention in responding to Toyota's requests. 2. Attempt to extract monopoly rents once the relationship is established.	1. Deliver parts late, or not to spec. 2. Fail to cooperate in the redesign of critical parts.
Toyota	1. Allow the supplier to make a "reasonable return" on their investments. 2. When things go wrong, provide resources to the supplier to help to fix them.	1. Fail to invest sufficient time or attention in responding to supplier's requests. 2. Attempt to extract monopoly rents once the relationship is established.	1. Cease working with the supplier.

Source: Derived from Helper and Levine (1992) and MacDuffie and Helper (1997).

different reasons. For example, Susan Helper made several visits to a GM parts plant (now closed) in Trenton (New Jersey) in 1990. Workers had been told that their suggestions would be welcomed and received training in Statistical Process Control. The workers responded by providing a flood of suggestions and by filling out charts tracking key quality metrics. However, management had not assigned anyone to respond to the suggestions, or examine the data the workers had carefully collected. The workers soon reverted to past patterns, feeling betrayed and much less interested in participating in future experiments.

Of course, it is difficult to disentangle the relative roles that credibility and clarity played in derailing General Motors. For instance, GM CEO Robert Stempel had been instrumental in trying to roll out the new practices across the company, but in 1991, in the depths of recession, GM North American operating losses reached nearly $5 billion, and Stempel was apparently unable to persuade either his board or Wall Street that these losses were not a sign of managerial "complacency"; so he switched gears, announcing plans to close 21 plants and eliminate 74,000 jobs. GM's credibility with its blue collar workforce was almost certainly severely damaged as a result. And it was further shaken when Stempel was fired and replaced by Jack Smith, formerly of GM Europe. Smith was one of the original negotiators of the NUMMI agreement but Ignacio Lopez, his new head of purchasing, behaved as if he placed no value at all on building relational contracts with GM's supply base, demanding that

Table 4
Cooperation, Defection, and Punishment within Heavyweight Product Development Teams at Toyota

	Action		
Agent	*Cooperate*	*Defect*	*Punish*
Team leader	1. Support team members in working together—ensure maximum possible support for the project. 2. When problems emerge, work rapidly and effectively to fix them. 3. Recognize each team member's contribution appropriately.	1. Fail to collect resources for the team, or to protect the team when threatened by the rest of the organization. 2. Play favorites with team members, and fail to reward them for their contributions.	1. Give negative reviews of team members to the rest of the organization, or refuse to work with them.
Team member	1. Bring all possible skills and capabilities to the team. 2. Invest in understanding the point of view and expertise of other team members. Be open to continuously learning. 3. "Go the extra mile" to ensure the success of the team, even when it fails to benefit oneself or one's home function.	1. Fail to invest sufficient time or attention in the team's work. 2. Drag one's heals in working with other members of the team. Put the interests of the home function first.	1. Disengage at work, or quit.

Source: Derived from Clark and Fujimoto (1991).

suppliers dramatically reduce their prices and sharing with rivals aspects of suppliers' designs that suppliers had thought would be kept confidential (Babson 1998). Thus, GM's early efforts to establish relational contracts (especially while failing to develop a coherent story about how long-term investments in such relationships would pay off) appeared unconvincing to outsiders. And unfortunately, the same financial losses in 1991 that finally convinced most of GM that it did need to change also removed some of the resources needed to make such a change possible.

It has only been within the last few years that General Motors has finally begun to rival Toyota's performance. The new "General Motors production system"—a GM-specific system that embodies many of Toyota's practices and includes stronger and more effective relational contracts—has been gradually diffused across the firm, largely as a result of the gradual diffusion across the firm of managers who had been trained at NUMMI and Saturn. The change came too late to save General Motors from bankruptcy.

Conclusions and Implications

We have suggested that General Motors first sharply declined and then failed for three reasons. First, it appears that GM's enormous success led it to deny the threat posed by foreign competition for over a decade. It is hard to be definitive as to whether this denial was due to the ossifying effect of decades of market power, of weak internal incentives, or of cognitive biases in the senior team, particularly since these factors surely interacted. One cannot even be sure that this denial was irrational, given the fact that many potential threats do not actually materialize and that the costs of significant change can be very substantial.

Second, even after General Motors recognized that its once-secure position was under threat, it took some time for the firm to understand the nature of the cluster of techniques that drove Japanese success. It had taken Toyota many years to develop what we now call "high performance work systems"—patterns of managerial practice that center around dense networks of communication and joint problem solving. Problems of complementarity and information localization may mean that the implementation of these techniques is subject to the informational constraints that limit the rate of diffusion of any complex new technology.

Third, these techniques could not be implemented without the simultaneous development of effective relational contracts, and General Motors had considerable difficulty building such contracts. GM's history of adversarial relations with its blue collar workers and suppliers, its reliance on an operating model that assumed there was very little need to exchange knowledge across either firm or functional boundaries, and its muddled and unaccountable internal processes made it difficult to build the credibility fundamental to effective long-term relationships. Moreover even after the firm had recognized the importance of adopting techniques like continuous improvement and cross-functional communication, it took some years to understand the nature of the relational contracts that would be needed to diffuse them within the organization. Managers accustomed to governing by fiat appear to have found it difficult to understand the potential benefits of discussion and collaboration. Even after the techniques had been widely diffused and problems of clarity seemed to have been overcome, problems of credibility continued to arise, as the collapse in supplier relationships following GM's apparent "defection" after its large financial losses in 1991 suggests.

If this cluster of explanations for the decline of General Motors is essentially correct, it raises some intriguing questions. To what degree are these explanations more broadly descriptive of the decline of so many of GM's contemporaries in the US manufacturing sector? What implications might they have for our understanding of management today? What are some potentially fruitful avenues for further research?

The decline of General Motors is exceptional in its scale, but many firms that dominated the US economy in the 1950s and 1960s echoed GM in responding slowly and ineffectively to the changing competitive landscape of the 1980s and

1990s. Global competition rooted in low-wage labor upended the apparel industry. New technology displaced Kodak and AT&T. Innovative domestic entrants replaced Sears and K-Mart. But at least two aspects of GM's experience seem common to a wide range of firms. First, past success often led to extended periods of denial. For example, the leaders of the American steel industry were extraordinarily slow to adopt competitive techniques (Christensen 1997), and most of the major semiconductor producers refused to believe that their Japanese competitors were outflanking them (Ferguson 1989). Indeed this pattern of denial following extended success appears to be a worldwide phenomenon. Nokia's recent collapse and Sony's decline suggest that it may be a danger for all large, successful firms. Even Toyota stumbled in 2010, in significant part due to its delayed reaction to customer quality concerns, although the automaker appears to have weathered the storm, regaining profits and market share since then. Second, many large American manufacturers had difficulty adopting the bundle of practices pioneered by firms like Toyota. One careful study of effects of human resource management practices on the productivity of integrated steel finishing lines, for example, suggested that "better" practices were surprisingly slow to diffuse: 36 percent of the monthly observations in the sample used the traditional, "command and control" bundle of practices for all five years of the study, and 58 percent still had no teams by the end of the period (Ichniowski, Shaw, and Prennushi 1997).

Some observers have suggested that these problems are evidence of "short-ism" in the American economy, or of an obsession with quantitative metrics, but the experience of General Motors suggests that this diagnosis is too simple. GM, particularly under Roger Smith in the early 1980s, made several very large investments—largely in technology—that were not expected to pay off for a long time. Instead, it appears that GM's management did not initially understand—or had difficulty communicating—the long-term value of relational contracts. Similarly Toyota appears to have collected more quantitative metrics than GM, and to have done so more frequently. It was GM's failure to use metrics as a guide to problem solving, rather than as a measure for internal control, that created difficulty.

Another stream of work suggests that the core problem is the nature of American labor and capital markets. For example, Hall and Soskice (2001) and others have argued that the institutional structure of the Japanese and German economies makes it relatively easier for firms in those countries to build long-term relationships with their workforce and their supplier base. However the success of the Japanese and German automotive firms in the United States—and GM's eventual adoption of Toyota's techniques—suggest that differences in institutional context alone cannot be determinative, although we cannot rule out the hypothesis that firms that have a history of relying on relational contracts in one market may have an advantage in building them in new geographies. It remains an open question how the kinds of institutions that appear to have contributed to Japanese and German manufacturing excellence might be adapted to help a US-owned firm remake itself by rewriting relational contracts and establishing new ones with workers and suppliers with whom it has a history of defection.

The potential advantages of relational contracts may have more widespread applications. Public support for economic growth has long focused on the diffusion of technology-based insights, but our work suggests that learning more about when (and what type of) relational contracts are likely to be valuable may be just as important. For example, many US firms appear to be managing their global supply chains on a spot basis, despite the fact that some intriguing research has suggested that this may be suboptimal on a number of dimensions (for example, Locke 2013). Further research exploring the conditions under which relational contracts may provide superior performance, and how such contracts can be built, may be particularly valuable in this context.

Viewing the firm through the lens of relational contracts may also have broader implications for the theory of the firm. In many economic models, the role of the manager is to make strategic choices—to decide which markets to enter and which inputs to combine—and to monitor subordinates. This view of the manager as a strategic architect or as an entrepreneur implementing a vision from the top down has deeply pervaded both the managerial literature and management practice. But although this perspective has proved illuminating and successful in many contexts, it omits the reality of building long-term relationships with people who often have very different interests and very different understandings of the world. Some suggestive empirical work, for example, indicates that "high commitment" or "purpose driven" firms[6] are particularly successful in motivating their workforce (Pfeffer 1998), perhaps because they are better positioned to communicate clearly and credibly both within and across firm boundaries and can thus build relational contracts more easily. Similarly some intriguing qualitative studies have suggested that being able to combine the ability to form relational contracts with the ability to hold everyone concerned to very high standards of accountability may be a particularly powerful skill (Beer 2009). If these findings are strengthened by ongoing research, the story of GM's decline has the potential to open some important doors for how researchers view the firm.

■ *The authors would like to thank the editors of this journal for their comments and suggestions, and would also like to thank Robert Gibbons, Casey Ichniowski, Mari Sako, Jeff Liker, Michael Wasser, Paul Adler, Brad Markell, and Jeffrey Morrow.*

[6] There is no precise definition of what it means to be either "high commitment" or a "purpose driven" firm, but in general these are firms in which the chief executive officer and his or her senior managers have made some kind of costly, public commitment to goals beyond short-term profit maximization.

References

Adler, Paul S. 1993. "The 'Learning Bureaucracy': New United Motor Manufacturing, Inc." In *Research in Organizational Behavior,* edited by Barry M. Staw and Larry L. Cummings. Greenwich, CT: JAI Press.

Adler, Paul S. and Bryan Borys. 1992. "Two Types of Bureaucracy: Enabling and Coercive." *Administrative Science Quarterly* 41(1): 61–89.

Adler, Paul S., Barbara Goldoftas, and David I. Levine. 1997. "Ergonomics, Employee Involvement, and the Toyota Production System: A Case Study of NUMMI's 1993 Model Introduction." *Industrial and Labor Relations Review* 50(3): 416–37.

Adler, Paul S., Thomas A. Kochan, John Paul MacDuffie, Frits K. Pil, and Saul Rubenstein. 1997. "United States: Variations on a Theme." In *After Lean Production: Evolving Employment Practices in the World Auto Industry*, edited by T. A. Kochan, R. D. Landsbury, and J. P. MacDuffie, 61–84. Ithaca: ILR Press.

Akerlof, George. 1984. "Gift Exchange and Efficiency-Wage Theory: Four Views." *American Economic Review* 72(2): 79–83.

Appelbaum, Eileen, Thomas Bailey, Peter Berg, and Arne Kalleberg. 2000. *Manufacturing Advantage: Why High-Performance Work Systems Pay Off.* Ithaca, NY: Cornell University Press.

Associated Press. 2007. "GM-UAW Back at Bargaining Table; Auto Stocks Rise." *USA Today*, September 17. http://usatoday30.usatoday.com/money/autos/2007-09-17-gm-uaw_N.htm?csp=34.

Babson, Steve. 1998. "Ambiguous Mandate: Lean Production and Labor Relations in the United States." In *Confronting Change: Auto Labor and Lean Production in North America*, edited by Huberto Juárez Núñez and Steve Babson, 23–50. Wayne State University Press.

Bachmann, Reinhard, and Akbar Zaheer, eds. 2006. *Handbook of Trust Research*. Northampton MA: Edward Elgar.

Beer, Michael. 2009. *High Commitment, High Performance: How to Build A Resilient Organization for Sustained Advantage.* Jossey Bass.

Behr, Peter. 1981. "A Search for Quality: Detroit Tries It All." *Washington Post*, August 23, p. G1.

Black, Sandra, and Lisa Lynch. 2001. "How to Compete: The Impact of Workplace Practices and Information Technology on Productivity." *Review of Economics and Statistics* 83(3): 434–45.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. 2013. "Does Management Matter: Evidence from India." *Quarterly Journal of Economics* 128(1): 1–51.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. 2013."Management as a Technology." Unpublished paper.

Board, Simon. 2011. "Relational Contracts and the Value of Loyalty." *American Economic Review* 101(7): 3349–67.

Brown, Clair, and Michael Reich. 1989. "When Does Union-Management Cooperation Work? A Look at NUMMI and GM-Van Nuys." *California Management Review*, vol. 31, no. 4, p. 26.

Chassang, Sylvain. 2010 "Building Routines: Learning, Cooperation, and the Dynamics of Incomplete Relational Contracts." *American Economic Review* 100(1): 448–65.

Christensen, Clayton. 1997. *The Innovator's Dilemma*. Harvard Business School Press. Boston, MA.

Clark, Kim, and Takahiro Fujimoto. 1991. *Product Development Performance: Strategy, Organization, and Management in the World Auto Industry.* Boston: Harvard Business School Press.

Cusumano, Michael A. 1985. *The Japanese Automobile Industry*. Cambridge, MA: Harvard University Press.

Cusumano, Michael A., and Akira Takeishi. 1991. "Supplier Relations and Management: A Survey of Japanese, Japanese-transplant, and U.S. Auto Plants." *Strategic Management Journal* 12(8): 563–88.

Dyer, Jeffrey H., and Nile W. Hatch. 2006. "Relation-specific Capabilities and Barriers to Knowledge Transfers: Creating Advantage through Network Relationships." *Strategic Management Journal* 27(8): 701–19.

Ferguson, Charles. 1989. "America's High-Tech Decline." *Foreign Policy* no. 74, pp. 123–44.

Finkelstein, Sydney. 2003. "GM and the Great Automation Solution." *Business Strategy Review* 14(3): 18–24.

Freyssenet, Michel, Andrew Mair, Koichi Shimizu, and Giuseppe Volpato. 1998. *One Best Way? Trajectories and Industrial Models of the World's Automobile Producers.* Oxford University Press.

Gibbons, Robert, and Rebecca Henderson. 2013. "What Do Managers Do? Exploring Persistent Performance Differences amongst Seemingly Similar Enterprises." Chap. 17 in *The Handbook of Organizational Economics*, edited by Robert Gibbons and John Roberts. Princeton University Press.

Halac, Marina. 2012. "Relational Contracts and the Value of Relationships." *American Economic Review* 102(2): 750–79.

Hall, Peter A., and David Soskice, ed. 2001. *Varieties of Capitalism: The Institutional Foundations of Comparative Advantage.* Oxford University Press.

Helper, Susan. 1987. "Supplier Relations and Technical Change: Theory and Application to the

US Automobile Industry." Doctoral Dissertation, Department of Economics, Harvard University.

Helper, Susan. 1991. "Strategy and Irreversibility in Supplier Relations: The Case of the US Automobile Industry." *Business History Review* 65(4): 781–824.

Helper, Susan R. 1994. "Three Steps Forward, Two Steps Back in Automotive Supplier Relations." *Technovation* 14(10): 633–40.

Helper, Susan. 2011. "Challenge and Opportunity in the U.S. Auto Industry: The Key Role of Suppliers." *Economia e Politica Industriale* 38(2): 51–67.

Helper, Susan, and David I. Levine. 1992. "Long-term Supplier Relations and Product Market Structure." *Journal of Law, Economics, and Organization* 8(3): 561–81.

Helper, Susan, and John Paul MacDuffie. 2008. "Better than a Bailout." *New Republic*, December 2. http://www.newrepublic.com/article/politics/better-bailout.

Helper, Susan R., and Mari Sako. 1995. "Supplier Relations in Japan and the United States: Are They Converging?" *Sloan Management Review,* Spring, pp. 77–84.

Huselid, Mark A. 1995. "The Impact of Human Resource Management Practices on Turnover, Productivity, and Corporate Financial Performance." *Academy of Management Journal* 38(2): 635–72.

Hyde, Justin. 2013. "First Pontiac Aztek's Sale Highlights the Long Half-life of Ugly." *Motoramic,* August 1.

Ichniowski, Casey, Kathryn Shaw, and Giovanni Prennushi. 1997. "The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines." *American Economic Review* 87(3): 291–313.

Ingrassia, Paul. 2010. *Crash Course: The American Automobile Industry's Road to Bankruptcy and Bailout—and Beyond.* New York: Random House.

Kaplan, Sarah, and Rebecca Henderson. 2005. "Inertia and Incentives: Bridging Organizational Economics and Organizational Theory." *Organization Science* 16(5): 509–521.

Keller, Maryann. 1989. *Rude Awakening: The Rise, Fall, and Struggle for Recovery of General Motors.* New York: Harper Perennial.

Kochan, Thomas A., Harry C. Katz, and Robert B. McKersie. 1986. *The Transformation of American Industrial Relations.* New York: Basic Books.

Levin, Jonathan. 2003. "Relational Incentive Contracts." *American Economic Review* 93(3): 835–57.

Li, Jin, and Niko Matouschek. 2013. "Managing Conflicts in Relational Contracts." *American Economic Review* 103(6): 2328–51.

Lieberman, Marvin B., and Shigeru Asaba. 1997. "Inventory Reduction and Productivity Growth: A Comparison of the Japanese and US Automotive Sectors." *Managerial and Decision Economics* 18(2): 73–85.

Lieberman, Marvin B., and Lieven Demeester. 1999. "Inventory Reduction and Productivity Growth: Linkages in the Japanese Automotive Industry." *Management Science* 45(4): 466–85.

Lieberman, Marvin B., Susan Helper, and Lieven Demeester. 1999. "The Empirical Determinants of Inventory Levels in High-Volume Manufacturing." *Production and Operations Management* 8(1): 44–55.

Liker, Jeffrey K. 2004. *The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer.* New York: McGraw-Hill.

Liker, Jeffrey K., and James Franz. 2011. *The Toyota Way to Continuous Improvement: Linking Strategy and Operational Excellence to Achieve Superior Performance.* New York: McGraw-Hill.

Locke, Richard M. 2013. *The Promise and Limits of Private Power Promoting Labor Standards in a Global Economy.* Cambridge University Press.

MacDuffie, John Paul. 1995. "Human Resource Bundles and Manufacturing Performance: Organizational Logic and Flexible Production Systems in the World Auto Industry." *Industrial and Labor Relations Review* 48(2): 197–221.

MacDuffie, John Paul. 1997. "The Road to 'Root Cause': Shop-Floor Problem-Solving at Three Auto Assembly Plants." *Management Science* 43(4): 479–502.

MacDuffie, John Paul, and Susan Helper. 1997. "Creating Lean Suppliers: The Honda Way." *California Management Review* 39(4): 118–51.

MacDuffie, John Paul, Kannan Sethuranman, and Marshall Fisher. 1996. "Product Variety and Manufacturing Performance: Evidence from the International Automotive Assembly Plant Study." *Management Science* 42(3): 350–69.

Milgrom, Paul, and John Roberts. 1990. "The Economics of Modern Manufacturing: Technology, Strategy, and Organizations." *American Economic Review* 80(3): 511–28.

Milgrom, Paul, and John Roberts. 1995. "Complementarities and Fit: Strategy, Structure, and Organizational Change in Manufacturing." *Journal of Accounting and Economics* 19(2–3): 179–208.

Nishiguchi, Toshihiro. 1994. *Strategic Industrial Sourcing: The Japanese Advantage.* New York: Oxford University Press.

Parker, Mike, and Jane Slaughter. 1994. "Working Smart: A Union Guide to Participation Programs and Reengineering." Detroit: Labor Notes Press.

Pfeffer, Jeffrey. 1998. *The Human Equation: Building Profits by Putting People First.* Boston: Harvard Business School Press.

Repenning, Nelson P., and Rebecca M. Henderson. 2010. "Making the Numbers? 'Short Termism' & the Puzzle of Only Occasional Disaster." NBER Working paper 16367.

Rubenstein, Saul A., and Thomas A. Kochan. 2001. *Learning from Saturn.* Ithaca, NY: Cornell University Press.

Russo, John. 1984. "Killing Jobs with 'Cooperation': The GM Memo." Labor Research Review 1(5): Article 10.

Sako, Mari. 2006. *Shifting Boundaries of the Firm—Japanese Company, Japanese Labor.* Oxford University Press.

Sloan, Alfred P. 1963, 1990. *My Years with General Motors.* Crown Books.

Smitka, Michael J. 1991. *Competitive Ties: Subcontracting in the Japanese Automotive Industry.* New York: Columbia University Press.

Soderberg, Leif G. 1989. "Facing Up to the Engineering Gap." *McKinsey Quarterly,* Spring, p. 3–23.

Sullivan, Mary W. 1998. "How Brand Names Affect the Demand for Twin Automobiles." *Journal of Marketing Research* 35(2): 154–65.

Takeishi, Akira. 2002. "Knowledge Partitioning in the Interfirm Division of Labor: The Case of Automotive Product Development." *Organization Science* 13(3): 321–38.

Taylor, Alex, III. 2010. *Sixty to Zero: An Inside Look at the Collapse of General Motors—and the Detroit Auto Industry.* Yale University Press.

"This American Life," National Public Radio. 2010. "NUMMI". Episode 403, March 26. http://www.thisamericanlife.org/radio-archives/episode/403/nummi.

Train, Kenneth E., and Clifford Winston. 2007. "Vehicle Choice Behavior and the Declining Market Share of U.S. Automakers." *International Economic Review* 48(4): 1469–96.

Weisman, Jonathan. 2005. "Biggest Automaker Needs Big Changes." *Washington Post,* June 11.

Welch, David, and Dan Beucke. 2005. "Why GM's Plan Won't Work." *Business Week,* May 8. http://www.businessweek.com/stories/2005-05-08/why-gms-plan-wont-work.

Womack, James, P. Daniel T. Jones, and Daniel Roos. 1990. *The Machine that Changed the World.* New York, NY: The Free Press.

Wright, J. Patrick. 1979. *On a Clear Day You Can See General Motors: John Z. DeLorean's Look Inside the Automotive Giant.* New York: Wright Enterprises.

Yates, Brock. 1983. *The Decline and Fall of the American Automobile Industry.* New York: Empire Books.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 73–98

Global Biofuels: Key to the Puzzle of Grain Market Behavior

Brian Wright

In the last half-decade, sharp jumps in the prices of wheat, rice, and corn, which furnish about two-thirds of the calorie requirements of mankind, have attracted worldwide attention. They have alarmed consumers, destabilized and even toppled some governments, and induced new temporary market distortions and bans.

These jumps in the prices of these three major grains have also revealed the chaotic state of economic analysis of agricultural commodity markets. Economists and scientists have engaged in a blame game, apportioning percentages of responsibility for price spikes to bewildering lists of factors, which include a surge in meat consumption caused by unprecedented increases of income of the vast populations of China and India; idiosyncratic regional droughts and fires; speculative bubbles; a new "financialization" of grain markets; the slowdown of global agricultural research spending; jumps in costs of energy and fertilizers; shifts in interest rates; the decline of the dollar; the surge in biofuels demands; bans on genetically modified plants; and climate change. Several observers have claimed to identify a "perfect storm" in the grain markets in 2007/2008, a confluence of some of the factors listed above.

The continuing confused state of the economics of grain price volatility may seem odd. After all, grain markets have many of the features of textbook competitive models. The products are relatively uniform. Their primary producers and ultimate consumers are atomistic price takers. Prices and outputs in the United States and other developed countries are unusually well measured by the standards of most goods and services in the world economy, with data freely available from

■ *Brian Wright is Professor of Agricultural and Resource Economics, University of California, Berkeley, California. His email address is bwright@berkeley.edu.*

http://dx.doi.org/10.1257/jep.28.1.73 doi=10.1257/jep.28.1.73

institutions including the US Department of Agriculture, the Food and Agriculture Organization of the United Nations, and the World Bank. Indeed, many pioneering empirical works in economics are related to agriculture, following the lead of the remarkable study of Ernst Engels (1821–1896) on the relation between expenditure on food consumption and income (for an overview in this journal, see Chai and Moneta 2010). Pioneers in econometrics estimated demand functions for agricultural products using random short-run disturbances in supply as a way of identifying a demand curve (Stock and Trebbi 2003). Other agricultural economists estimated early dynamic models of supply response (Nerlove 1956) and of technology change and diffusion in agriculture (Griliches 1957). Given the longstanding empirical attention paid to agricultural markets and the many attractive features of commodity market data, why is there so little clarity on the causes of recent fluctuations in grain prices?

In this paper, I will show, using data on price and production, that the most basic Marshallian model of agricultural supply and demand, in which annual production is all consumed in the same year, does not explain major annual movements of the three major grains.[1] However, the Marshallian model explains price movements surprisingly well until 2004 *if* extended to recognize two key kinds of substitution: the substitution of one of the major grains for another (for example, wheat substituted for corn or rice consumed by humans or for corn consumed by farm animals); and the substitution, via storage, of grain harvested in one year for grain harvested and consumed in later years. Until 2004, storage and intergrain substitution were the two essential keys to understanding the economics of grain market behavior in a Marshallian microeconomic model. However, the kind of intertemporal price smoothing via storage observed up to 2004 is not evident in price and storage behavior since then. In particular, even when different types of grains are aggregated, large increases in price occurred in years when stocks carried over to the next year were also increasing. Many economists, puzzled by this phenomenon, concluded that competitive microeconomic models could not explain recent market behavior. Those who searched farther afield came up with confusing lists of causative factors, ranging from yield effects of global warming, to income surges in China and India, to financial speculation.

There was also great confusion as to the distributional effects of the huge grain price increases. Important institutions that had recently argued that low food prices exacerbated food insecurity and poverty by cutting incomes of farmers in low-income countries now issued reports that the sharp rise in food prices had worsened the plight of the hungry and poor.[2] The Renewable Fuels Association in 2011

[1] Soybeans and rapeseed (including canola) are two major oilseeds (not generally included in "grains") that compete with the grains for land, and are inputs for diesel biofuels, and are important substitutes with grains as sources of calories for humans and farm animals. Although they are important, for tractability, here I limit the discussion to the three major grains.

[2] Swinnen and Sqiucciarini (2012) note that in statements three years apart the Food and Agriculture Organization claimed that the falling price trend "threatens the food security of hundreds of millions of people," and that "rising food prices . . . worsen the food deprivation suffered by 854 million

denied claims that corn ethanol production caused high food prices, but declared in a November 15, 2013 news release that reducing the requirements for ethanol production would lead to sharply lower prices for farmers.[3]

The price jumps since 2005 are best explained by the new policies causing a sustained surge in demand for biofuels. The resulting reduction in available per capita supply of food and animal feed could not be accommodated by drawing on available stocks, as they had in the past when there were temporary shortages created by yield shocks. Instead, the necessary adjustments included an expansion of global net acres planted to grains, especially in Latin America and the former Soviet Union, and by reduced per capita consumption of grains and products from animals fed on grains. (There was no noticeable increase in crop yields from trend in the United States.) Thus to solve the puzzle of recent grain market behavior it is necessary to incorporate into the market model—in addition to substitution between grains as sources of calories, and substitution between successive harvests via storage—a third key substitution, that of biofuels for petroleum-based fuels.

The rises in food prices since 2004 have generated huge wealth transfers to global landholders, agricultural input suppliers, and biofuels producers. The losers have been net consumers of food, including large numbers of the world's poorest peoples. The cause of this large global redistribution was no perfect storm. Far from being a natural catastrophe, it was the result of new policies to allow and require increased use of grain and oilseed for production of biofuels. Leading this trend were the wealthy countries, initially misinformed about the true global environmental and distributional implications.

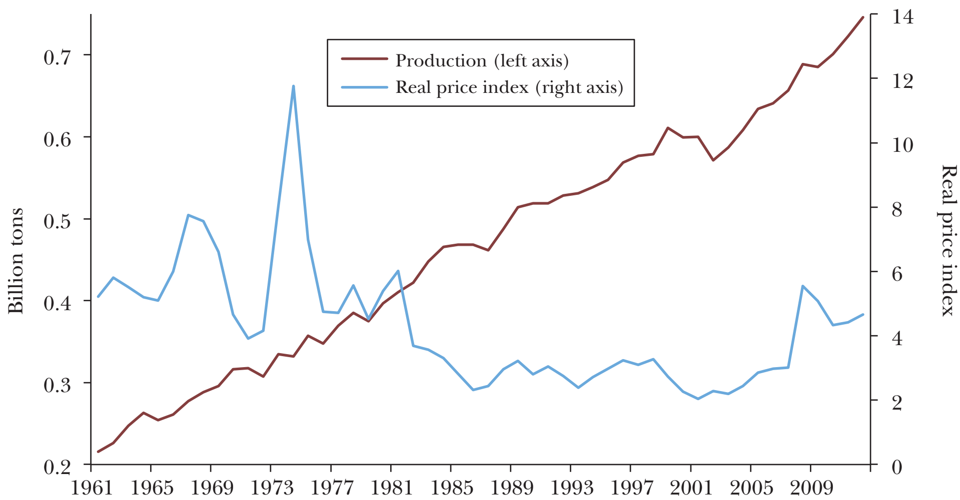
Grain Market Behavior through 2004

In this section, I discuss the behavior of grain markets from 1961 through 2004, identifying 2005 as the first year of a new market regime. I begin with a simple supply-demand model in which production is seasonal with one grain harvest per year. Over time, the yield has a positive trend due to persistent productivity increases.

people." Similarly Oxfam, which had previously claimed that as a consequence of low prices "over 900 millions of farmers are losing their livelihoods," later declared that high food prices had pushed millions of people in developing countries "further into hunger and poverty." A February 2008 US Department of Agriculture report declared "Strong expansion of corn-based ethanol production in the United States affects virtually every aspect of the field crops sector, ranging from domestic demand and exports to prices and the allocation of acreage among crops. A higher portion of overall plantings is allocated to corn. Higher feed costs also affect the livestock sector . . ." (p. 5). Five months later, in a letter to Jeff Bingaman, Chair of the Senate Committee on Energy and Natural Resources, the Secretary of Agriculture with the Secretary of Energy, after endorsing Bingaman's statement that the issues were complex, cautioned against "hasty judgments driven by highly questionable agenda-driven calculations," then went on to state: "Our preliminary analysis . . . suggests that current biofuels-related feedstock demand plays only a small in global food supply and pricing."

[3] http://www.ethanolcfa.org/news/entry/the-epas-rvo-proposal-cannot-stand (last accessed January 12, 2014).

Figure 1
Rice: Real Price Index and Production, 1961–2012



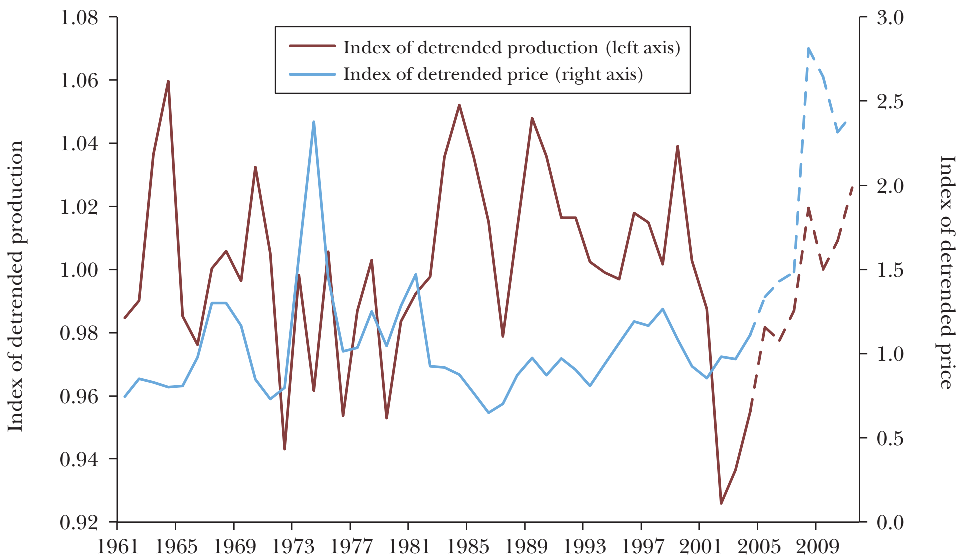
Notes: The nominal rice price from 1961 to 2012 is the calendar year annual average price for rice (Thailand 5% broken) from the commodity price dataset of the World Bank (the pink sheet). The real price index for rice is obtained by deflating the nominal price using Manufactures Unit Value (MUV) from the same dataset. The world rice production from 1961 to 2011 is the calendar year production data for rice paddy from the FAOSTAT of the UN Food and Agriculture Organization. For 2012, we use the US Department of Agriculture (USDA) production data for rice (milled) for the marketing year 2011/2012 adjusted by a ratio between the FAOSTAT 2011 data and USDA 2010/2011 data to roughly account for the difference between paddy and milled rice.

At planting time, the anticipated harvest is roughly proportional to planted area, but the realized harvest is subject to roughly proportional random disturbances, including weather fluctuations, pest infestations, and other shocks that have shown little evidence of inter-year persistence. Grain calorie consumption falls as prices rise, and has an upward trend over time due to (exogenous) population increase; I mainly ignore income and other shifters as unimportant in the short run. As a starting point, can this straightforward economic approach make sense of outcomes in the market for any of the three major grains?

Take rice, for example. As Figure 1 shows, world rice production has been following a remarkably linear upward trend, reflecting strong productivity improvements with varying but modest percentage deviations from that trend. Between 1961 and 2004, prices (deflated by the Manufactures Unit Value, which removes the direct effect of changes in the value of the dollar) followed a downward trend, interspersed with intervals of large variation characterized by price spikes—that is, sharp jumps soon followed by similarly sharp reversions toward trend.

If rice demand evolves steadily, then according to this simple model a price spike must be associated with a shift in inelastic short-run supply. However, as

Figure 2
Rice: World Detrended Price versus Index of Detrended Production, 1961–2012



Notes: The real price index is assumed to be trend stationary. The index of detrended price is obtained by taking the exponential of the residuals from regressing log real price against a constant and time. The world production is assumed to be the product of a linear function of time and a stationary shock. We regressed the log production against the log of the linear function time. We take the exponential of the resulting residuals and use the series as the index of detrended production. The correlation coefficient between the resulting two detrended series through the year of 2004 is –0.2. The dashed lines indicate a new market regime in effect after 2005.

Figure 2 shows, price spikes have not been accompanied by notable year-on-year production downturns. The largest fall in production by far, about 7 percent of trend production, occurred in 2002, a year when price rose only slightly. The simple supply and demand model clearly fails to explain behavior of rice markets during those periods in which the price displays a spike. For rice, the oft-heard argument that the workhorse supply-and-demand model does not explain grain price jumps is well-supported by the evidence—and this is true even before more recent price changes generated further doubts about the conventional market model.

Adding Storage to the Marshallian Market Model

Does the supply and demand model perform better if extended to recognize one key characteristic of a crop such as rice: grain harvested one year can, via storage, be substituted in future consumption for grain harvested later? Assume that grains can be stored from period to period, while acknowledging the reality that grain stocks cannot be negative. For simplicity ignore any cost, waste, or "shrinkage" associated with storage activity apart from a constant opportunity cost of capital, and assume planned production remains fixed at some positive level. Storers maximize expected profits, competing with other storers as well as consumers for the available supply, if the return expected from selling their grain at a higher price in the future covers the interest on their investment in stocks.

In this extended model, the available supply of a product in any year is the sum of the current harvest and any stocks carried over from the supply available in the previous year.[4] Similarly, total demand in any given year will be the amount consumed in that year and the amount that goes into storage for next year. This would all be very simple, but for one problem: How can storers' demand for stocks be determined?

The expected return to storage will vary depending on the available supply in the current period and on the market demand in the next period. If the storage demand curve in the next period were known, horizontally adding this demand curve to the known consumption demand would result in the market demand for the next period. Then we could solve for current storage demand. Unfortunately, we do not in general know the storage demand function in the next period.

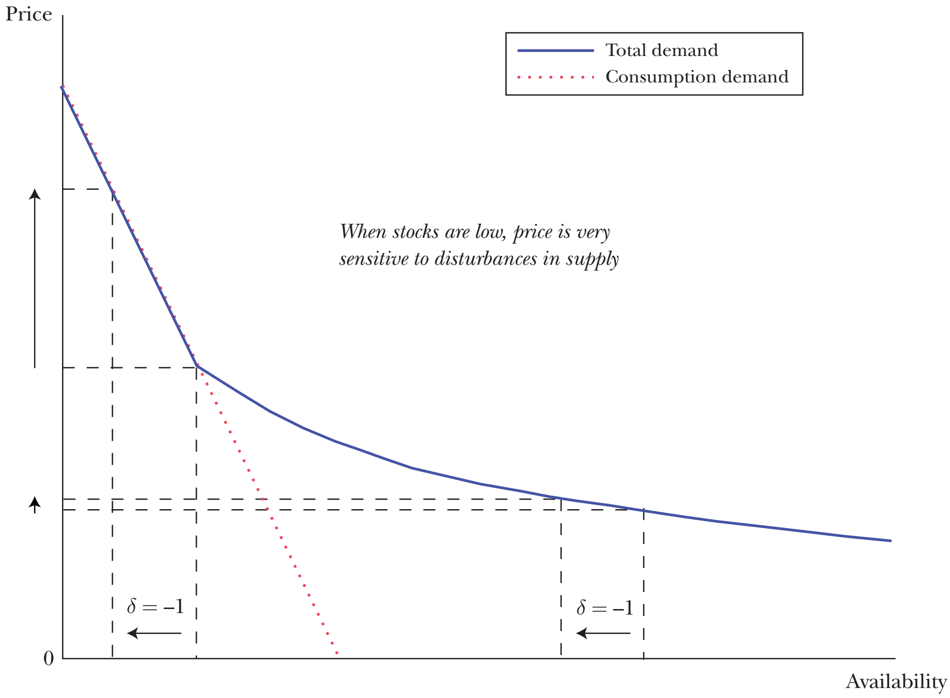
The first numerical model for solving for the storage demand was presented in the remarkable paper of Gustafson (1958), assuming what later became known as rational expectations (Muth 1961). Assuming a far future period beyond which storage is not possible, storage demand is solved by backward induction, exploiting the effects of discounting, that is, by dynamic programming (Bellman 1958). Readers unfamiliar with the topic can find what they need to understand the method as applied to a grain market in the simple exposition of Gardner (1979), or in Williams and Wright (1991).

In this framework, assuming no wastage, the market demand for grain shown in Figure 3 is the horizontal sum of the consumption demand and the storage demand. On the horizontal axis, availability refers both to the sum of current production and stocks carried over from the previous year ("carry-in stocks"). Consumption demand is the downward-sloping straight line, which in the figure is pictured as solid at the top and dashed at the bottom. When the current price is low, storers hold stocks to carry over to the following year. Thus, the curved area of the demand curve to the right of the kink represents the addition of "carry-out stocks" to consumption demand.

The figure illustrates that the effect of a transient exogenous harvest shock depends crucially on the initial available supply. Assume for simplicity that expected production is fixed. Assume further that carry-in stocks are so high that, if the harvest is at its mean, available supply is 7 units. If the realized harvest turns out to be one unit lower, almost one unit of stocks is consumed as a substitute for the missing production, so price need rise only a little to induce the small reduction

[4] For the purposes of this paper, stocks are defined as discretionary stocks net of the "pipeline" or "working" stocks essential for the operations of the market including, for example, grain in transit and grain needed for operation of marketing, processing or feeding activities. We also ignore a small amount of stocks that might be diverted to consumption but only at a very high and fast-increasing marginal cost, that is, stocks that might be said to have high "convenience yield."

Figure 3
Differential Price Responses to a Temporary Yield Shock

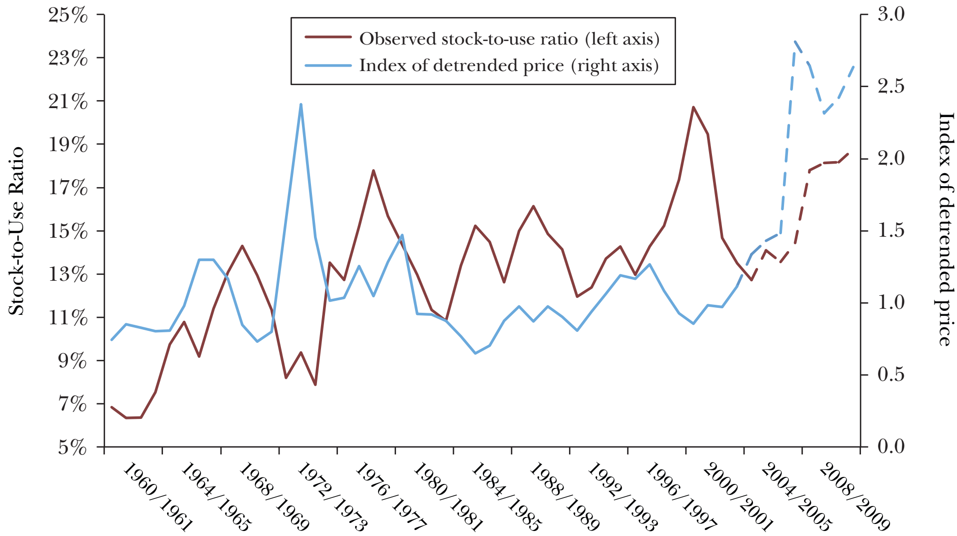


Notes: Price is the price of the grain. On the horizontal axis, availability refers both to the sum of current production and stocks carried over from the previous year. The figure illustrates that the price effect of a transient exogenous harvest shock depends crucially on the initial available supply.

in consumption needed to absorb the fraction of adjustment not covered by stock reduction. If carry-in stocks are so low that adding mean production would place available supply at the kink in demand, then the price rise must be far larger to induce the one unit drop in current consumption necessary given no mean output increase.

Figure 3 shows that in this model in which shocks come from temporary supply-side harvest disturbances, price falls when stocks rise, and price is negatively related to the ratio of stocks to consumption or "use." Figure 4 shows detrended real global rice price (as in the earlier Figure 2) as well as the observed stocks-to-use ratio. The numerator of the ratio is observed stocks, including slow-changing essential "pipeline stocks" as well as the discretionary stocks discussed. The denominator approximates consumption as "use," typically calculated as production less changes in stocks. We first estimated a trend on the sample truncated at 2004, to avoid the influence of a possible change in market regime after that year. The correlation between detrended rice price and the stock-to-use ratio is negative, –0.1355, even smaller in magnitude than the relation between rice price and current production.

Figure 4
Rice: Detrended Price versus Observed Stock-to-Use Ratio



Source: The stocks and consumption (use) data are from the Production, Supply, and Distribution (PSD) Online of the US Department of Agriculture.
Notes: In the stocks-to-use ratio, the numerator is observed stocks, including slow-changing essential "pipeline stocks" as well as the discretionary stocks discussed, while the denominator approximates consumption as "use," typically calculated as production less changes in stocks. The stocks-to-use ratio excludes Chinese stocks and use. The index of detrended price for rice is the same as in Figure 2. The trend is estimated through 2004, to avoid the influence of a possible change in market regime after that year. The correlation coefficient between the two series till 2004 is –0.1355.

Note also that the stock-to-use ratio is clearly trending upwards and positively related to price for several successive years early in the sample interval, and for the late 1990s. Apparently, adding the possibility of intertemporal substitution of rice via storage does not render the basic supply-and-demand model capable of consistently rationalizing fluctuations in the price of rice.[5]

The Importance of Substitution between Grains

In most parts of the world, wheat, rice, or corn is the strongly preferred staple food, the others being, for many consumers, poor short-run substitutes. However, in some regions including substantial parts of the vast Indian and Chinese populations, both wheat and rice figure prominently in the consumption basket. Further, as incomes increase over time, wheat is at least partially displacing rice, corn, and other staple human calorie sources such as coarse grains or tubers.

[5] Economic observers were understandably puzzled by this behavior, and looked beyond Marshallian models to explain the rice price spike of 2007/08. For example, Heady and Fan (2010, p. xiii) conclude, "The surge in rice prices stands apart as being almost entirely a bubble phenomenon."

Nonfood demands are important for corn and wheat. In many countries, an important part of the wheat supply goes to animal feed, and in Europe, wheat is also a significant input into biofuels production. Although corn is an important staple in some parts of Africa and South America, it is much more significant globally as an animal feed. Animal feed has a much higher global price- and income-elasticity of demand than does grain for human consumption. In the United States, corn is the dominant input for the grain ethanol industry. In some areas of the United States, corn competes with wheat for land (and, in the corn belt, with soy). Corn yields in the United States are often higher if corn is grown in rotation with soybeans. In parts of India, China, and other countries, wheat competes with rice for land, often in different multicrop rotations that might include both. The three major grains also compete for inputs such as fertilizer and water.

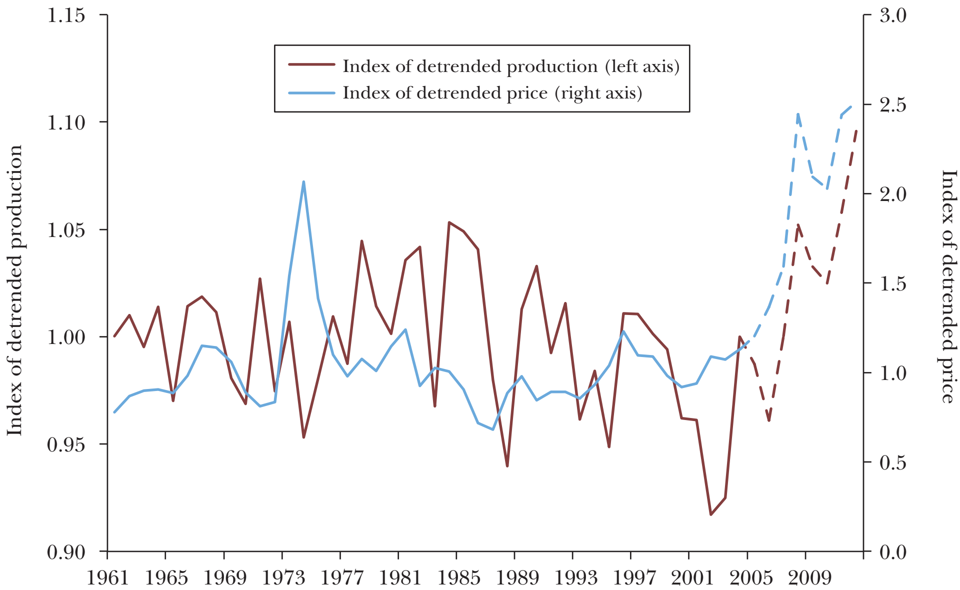
Is the substitutability between the three grains strong enough to more reasonably consider them as close substitutes rather than as essentially independent in consumption? To explore this question, consider the market for the three grains together as a market for aggregate calories, following the initiative of Roberts and Schlenker (2009), who also include soybeans, which has more distinct and higher-value markets in meal and in oil. The price of calories is constructed as the average of the annual prices of wheat, corn, and rice, with the weights being the world production of calories from each grain.[6] They found that the detrended price of calories for all grains taken together is highly correlated (at least 0.93) with the detrended real price of each grain over the years 1961–2012. Although on average the grains are consumed in ways that are quite distinct, on the margin they appear to be quite substitutable. In this aggregated market, does the supply and demand analysis perform better than it did for the rice market alone?

Figure 5 shows that the answer is no. The two largest production shortfalls from trend—measured in terms of calories produced—do not coincide with prominent price spikes. As for rice alone, production variation alone cannot explain prominent features of calorie price behavior, even prior to the onset of increased volatility after 2004.

Combine intertemporal *and* intergrain substitution possibilities in one model and the picture changes completely. Figure 6 shows that when storage of the combined major grains is considered, price behavior of aggregate calories from

[6] World wheat, maize (corn), and rice nominal price data are from World Bank/GEM Commodities. Wheat is measured as Wheat U.S. Hard Red Winter, maize is no. 2 Maize, and rice is measured as Rice Thailand 5%. The annual price is the monthly price observed in the last month of the marketing year: the wheat annual price is the May price, the maize annual price is the August price, and the rice annual price is the July price. All annual price data are deflated into real price indices using the annual Manufactures Unit Value Index from World Bank/GEM Commodities, which is a composite index of prices for manufactured exports from the 15 major developed and emerging economies to low- and middle-income economies, valued in US dollars. This index behaves very differently from the United States Consumer Price Index, especially in recent decades. Production data for the weights come from US Department of Agriculture (USDA), Foreign Agriculture Service (FAS), Production, Supply, and Distribution Online (PSDO). The weight-calories conversion rates are from USDA National Nutrient Database.

Figure 5
Calories: Index of World Detrended Price versus Index of World Detrended Production

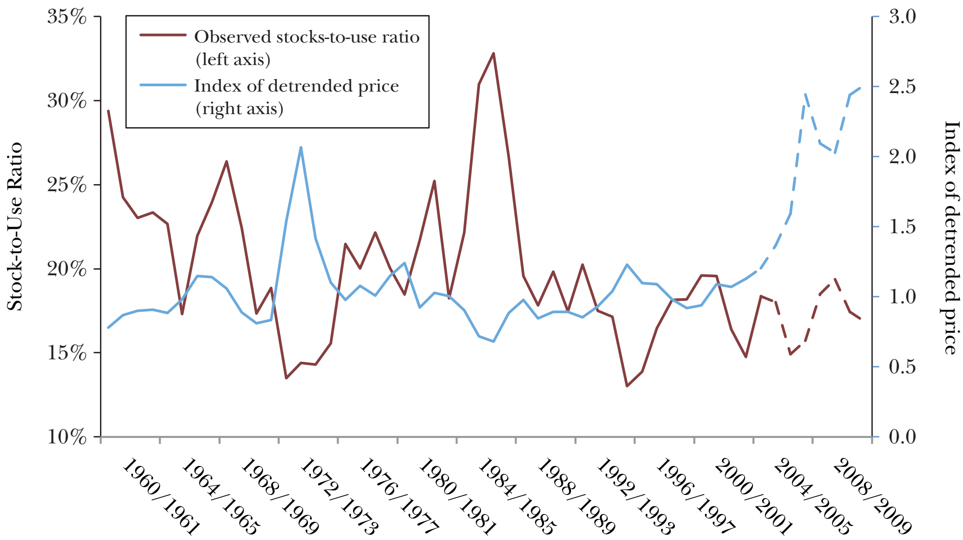


Notes: We convert the production quantities for corn, rice, and wheat to calorie units using the calorie content data from the US Department of Agriculture (USDA) National Nutrient Database. The world grains calorie production is obtained by aggregating calorie contents in the production of these three grains. The nominal price for calories is obtained by dividing the total nominal value of grain calorie production by total grain calorie production. After deflation, we obtain the detrended real price, which is detrended using the same technique as outlined in Figure 2. The world production of the three grains from 1961 to 2011 is the calendar year production data from the FAOSTAT of the UN Food and Agriculture Organization. For 2012, we use the USDA production data for corn, rice (milled), and wheat for the marketing year 2011/2012 using the technique described in the note to Figure 1 to generate the rice production data in 2012. The dashed lines indicate a new market regime in effect after 2005.

the major grains up to 2004 becomes highly intuitive. Each of the major price spikes in Figure 6 through 2004 is matched by a low observed value of the stock-to-use ratio, and the simple correlation between the two series is –0.5645.

Thus the reality of substitution between rice and the two other major grains is crucial to understanding the role of storage in stabilizing random and independent annual disturbances in production, and the role of storage is crucial to understanding intergrain substitution. Attempts to explain spikes in agricultural prices by looking only at annual production, or only at stocks of that specific grain, are futile. Rice price was rising in the early 1990s even though rice stocks were rising, because aggregate grain calorie stocks were falling, the three grains are substitutes in consumption, and aggregate stocks were falling. Likewise, the mystery of a spike in the rice price in 1973, when rice production was normal, is explained

Figure 6
Calories: Index of Detrended Price versus Observed Stock-to-Use Ratio



Notes: In the stocks-to-use ratio, the numerator is observed stocks, including slow-changing essential "pipeline stocks" as well as the discretionary stocks discussed, while the denominator approximates consumption as "use," typically calculated as production less changes in stocks. The stocks-to-use ratio excludes Chinese stocks and use. The stocks and consumption (use) data for corn, rice, and wheat are from the Production, Supply, and Distribution (PSD) Online of the US Department of Agriculture. Grain calorie consumption and stocks are constructed as for the grains calorie production in Figure 5. The index of detrended price for calories is the same as in Figure 5. The correlation coefficient between the two series through 2004 is –0.5645. The dotted lines indicate a new market regime in effect after 2005.

by substitution with the other grains and a shortage in available aggregate calories supply due to calorie stocks, combined with fairly low harvests of wheat and corn, relative to trend. Using the logic from Figure 3 earlier, low stocks make markets vulnerable to temporary harvest shortfalls.

Most empirical estimates of demand for grains and other foods ignore the distinction between stocks and consumption.[7] They typically fit or assume a market demand curve that ignores the kink shown in Figure 3 and find a slope or elasticity that averages the response of price to supply shocks above and below the kink. When discretionary stocks are low, such a smoothed market demand underestimates the response of price to a shock in available supply. The result is that if a global supply shock hits at a time when stocks are low, the resulting price jump is much higher than predicted by the model. This limitation is a common feature of computable general equilibrium models addressing policy analysis for food and agriculture,

[7] A recent example is Roberts and Schlenker (2013), which uses storage effects creatively to identify supply response, but does not model storage demand or the associated nonlinearity of market demand.

including measures to address climate change. Without a proper model of storage demand, they are incapable of assessing the effects of pre-announced, sustained increases in mandated use of grains for biofuels, discussed below.

These shortcomings of existing models are understandable, given that until recently there was no satisfactory way to distinguish empirically the kinked market demand from the consumption demand in estimation of commodity price responses to shocks. In particular, the pioneering empirical applications of the model outlined above by Deaton and Laroque (1995, 1996) to several commodities failed to replicate observed price correlations. However, in Cafiero, Bobenrieth, Bobenreith, and Wright (2011), using a more accurate numerical procedure in a model that otherwise replicated the approach of Deaton and Laroque, we obtained results consistent with observed price correlations for several commodities, including corn.

Building on this work, in Bobenrieth, Bobenreith, and Wright (forthcoming), we extended the model to include trends in prices and production while allowing consistent maximum likelihood estimation, following the approach in Cafiero, Bobenrieth, Bobenrieth, and Wright (2013). We estimated this model for each of the major grains and for aggregate grain calories, again using only global price data, up to the year 2007. The results show that a surprisingly large portion of the variation in a grain calorie price index and in stocks of grain calories prior to 2005 is explained by a simple Gustafson-style model that includes stocks, intergrain substitution, and expected profit-maximizing intertemporal arbitrage. The results also show that for the time period up to 2004 there is little reason to resort to "financialization" of the grain futures markets, irrational herding or speculation, or drops in the cost of capital to explain price spikes.

Why Has Grain Price Behavior Changed Since 2005?

Since about 2005, international grain markets have seen several of the largest price jumps since the 1970s. Careful readers of Figures 4 and 6 might also have noticed that before 2005, when prices rose, stocks typically fell, but after 2005 the relation between changes in prices and changes in stocks became positive, both for calories and for rice. Why would higher stocks of grain be correlated with higher prices for grain?

Before considering alternate explanations, I start with what I consider to be the key to this new puzzling behavior: the surge in demand for grain and oilseeds to produce biofuels.

Biofuels

The surge in biofuels production was driven by policies, led by the European Union and the United States, that allowed increased maximum shares of biofuels in blends with gasoline or diesel fuels, accompanied by mandated minimum use of biofuels for transport fuels, and supplemented in several cases by subsidies and/or import tariff policies. In the United States, as early as 1978, the US Energy Tax Act established tax credits for ethanol blenders. In 1990 the Clean Air Act Amendments mandated the use of either a refining byproduct like MTBE (methyl tertiary butyl ether) or ethanol as a gasoline "oxygenator" that would act to lower emissions of carbon monoxide. MTBE was cheaper than ethanol, and so was widely adopted. However, concerns arose that MTBE might be carcinogenic when gasoline carrying it leaked from old gas station tanks, polluting groundwater. In 1999 the California government banned use of MTBE, effective as of 2003; by 2006, 25 other states had also banned its use. As a result, there was a boost in demand for ethanol as a substitute for MTBE circa 2005. In addition, the 2005 Energy Policy Act introduced biofuels mandates as a policy instrument, in the form of 4 billion gallons of renewable transport fuels in 2006, rising to 7.5 billion gallons by 2012. In 2007 the Energy Security and Independence Act mandated an increase to 15 billion gallons by 2015. Moreover, increasing amounts of advanced biofuels—defined to include cellulosic ethanol and ethanol from Brazilian sugar cane, but not corn ethanol—were added to the annual mandate, rising to 21 billion gallons by 2022.

In the European Union, some limited support for biofuels crops began in 1988, but all such support ended in 2010 (Amezaga, Boyes, and Harrison 2010). Meanwhile, a 2003 European Union directive set biofuels targets of 2 percent of transport fuels for 2005 rising to 5.75 percent by 2010. Many countries chose to use tax incentives, allowed in another 2003 directive, to increase biofuels use. As tax incentives began to take effect, the resulting revenue losses induced states to shift to quantitative measures. Although biofuels use doubled between 2003 and 2005, the actual 2005 share reached only 1.4 percent of European Union transportation fuels (Amezaga, Boyes, and Harrison 2010). In 2009, as part of a climate and energy policy package, there was a shift to European Union use of mandates that included a 20 percent share of renewables in total energy consumption and a 10 percent share of biofuels in gasoline and diesel. Numerous other countries reportedly have enacted mandates for biodiesel or bioethanol blends.[8] World fuel ethanol demand increased from 7.5 billion gallons in 2005 to 22.7 billion gallons in 2012. World biofuels production increased from 14.7 million tons oil-equivalent in 2003 to 60.2 million tons in 2012, of which the United States and the European Union produced 27.4 and 9.9 million tons respectively.

By the standards of agricultural policy changes, the introduction of grain and oilseeds biofuels for use in transport fuels was abrupt, and the effects on the balance between supply and demand were dramatic. After 2003/04, corn used for ethanol in the United States doubled to two billion bushels in just two years, and doubled again in the following few years. In 2001, the projection from the Agricultural Baseline Database of the US Department of Agriculture was that the use of corn for fuel would rise from about 800 million bushels in 2001 to 1.1 billion bushels by 2013. However, by 2004, the actual use of corn for fuel had already risen

[8] These reportedly include Canada, China, Colombia, Costa Rica, Ethiopia, India, Jamaica, Kenya, Malawi, Malaysia, Panama, Paraguay, Peru, the Philippines, South Korea, Taiwan, Thailand, and Uruguay. New Zealand and South Africa have considered and rejected mandates (*Biofuels Digest* 2011).

to 1.9 billion bushels and was projected to reach 2.8 billion bushels by 2013. By 2009, 4.0 billion bushels of corn were being used for fuel, and the projection was 4.6 billion bushels by 2013.

Currently, biofuels use accounts for about one-third of United States corn production, net of byproducts used as animal feed. Further rise is halted by the "blend wall," reflecting the fact that much of the automobile fleet and the distribution system are not currently capable of handling a gasoline blend greater than the 10 percent ethanol (E10) fuel specified in current regulations.

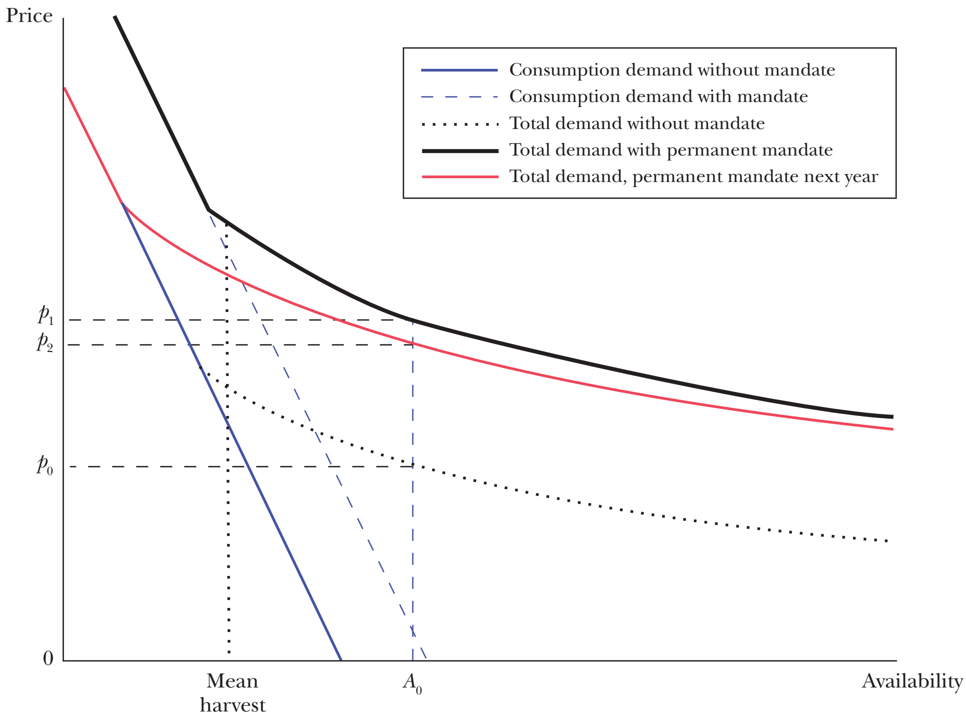
The effect of a given mandate on the evolution of prices depends crucially on whether the mandate is temporary or permanent. In the case of unanticipated imposition of a transitory (one-time) mandate for diversion of one unit from the market, the fall in consumption induced by this price rise absorbs only a small fraction of the shift; the temporary shortage is buffered by reducing storage to the next period, when the mandate will be over. In the next period, the market demand curve reverts to its original kinked form, and part of the reduction in supply will in turn be shared with future periods if the harvest is large enough to allow carry-out storage.

When initial stocks are larger than the mandate, arbitrage buffers the immediate effect of a temporary mandate just as it buffers a supply shock, by reducing carry-out stocks as illustrated in Figure 3. Note that if initial available supply is so low that no stocks are held, the immediate price rise from the mandate is far larger: the vertical distance between the new and old consumption demand curves at that supply level. In that case, the immediate price jump must be sufficient to reduce food and feed consumption enough to fully accommodate the mandate. Future prices are unaffected.

Consider now the case of the surprising and immediately-effective imposition of a mandate known to be permanent at its new level, illustrated in Figure 7. The opportunity to cushion the shock is limited by the fact that, by assumption, at any given price, the mandate increases total consumption in every future period. The outermost, thick market demand curve in Figure 7 reflects the fact that future demand curves remain at the same elevated level. The dynamics of similar permanent shifts in the market were introduced in Wright and Williams (1984). At the initial available supply A_0, upon news of the mandate price jumps to p_1, much higher than p_0. In contrast to the case of an equivalent but temporary mandate, food and/or feed consumption must immediately absorb most of the initial unit increase in mandated consumption. Depending on the details of the specification of the model, there is little or no buffering from stock adjustment. Even if there are substantial initial stocks, the slope of the consumption demand is a better guide to the size of the price jump induced by the mandate than is the slope of market demand including storage demand.

If more realistically, a permanent mandate is implemented one period after its announcement, and supply at announcement is as before, Figure 7 shows the market demand upon announcement as the second-highest, thinner demand curve. Price rises a little less, to p_2 but carry-out storage rises to a level higher than if the mandate were immediately imposed. Because the initial available supply is sufficient

Figure 7
Demand Shifts and Price Jumps upon Announcement of a Permanent Mandate



Notes: $p_2 - p_0$ represents the price rise due to anticipation of a permanent mandate next year.

to allow carryout storage, price and stocks both jump upwards in anticipation of the future, mandated demand boosts. If planned production responds (with a one period lag) to higher prices, then observed supply is also likely to increase the year after announcement.

Simultaneous jumps in prices, supply, and stocks are not observed in the behavior of storage in a market, such as the market for grain calories before 2005, in which the dominant disturbances are one-year supply disruptions. It is understandable that market observers were puzzled when prices, production, and stocks of grain calories all rose sharply between 2006 and 2008, as Figures 5 and 6 show.

This discussion of dynamics implies that sudden mandated upshifts in the rising paths of biofuels requirements for corn ethanol should have caused sharp jumps in stocks of grains at given use levels and sharp rises in price at given stocks levels. Indeed, Figures 5 and 6 (shown earlier) show simultaneous rises in price, production, and stocks of calories between 2006 and 2008, consistent with a strong shift in market demand due to anticipation of the effects of upward revision in the path of future mandates.[9] This analysis suggests that there is no need to look to "speculation" to explain episodes of increases in stocks accompanied by jumps in prices. This analysis is also consistent with the early conclusion of Mitchell (2008) that biofuels policy was the major driver of the price spikes he had recently observed and with the prior predictions of Runge and Senauer (2007) that biofuels expansion would have such effects on grain markets.

The "Asian Income Shock" Hypothesis

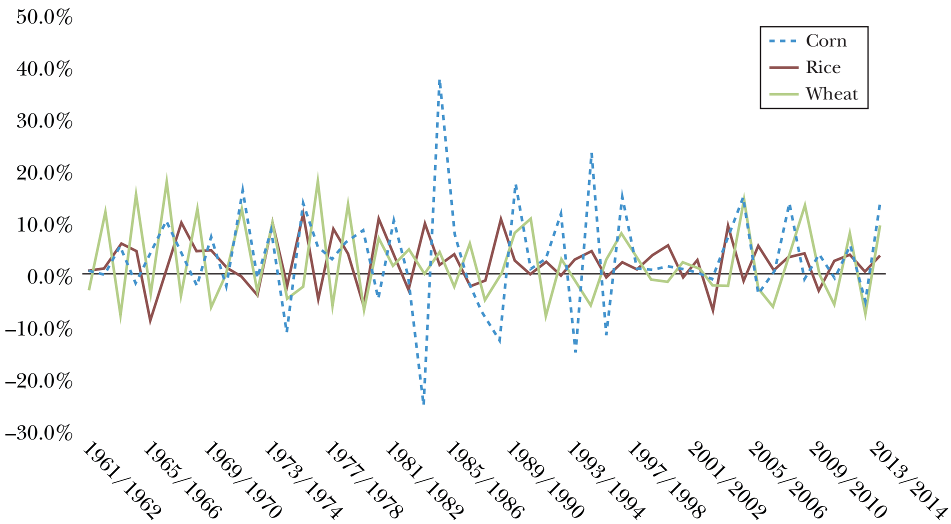
In the literature on recent grain price increases, another possible explanation points to the surge in income in India and China (for example, Krugman 2008; Brown 2012). Historically, one would expect any price shift due to national income changes to be modest; after all, annual per capita income changes in a given population are usually of the order of 4 percent in very good years, and food is income-inelastic. However, per capita incomes in both China and India have recently been growing at unprecedented rates, and their populations are large enough that income-induced variation in their consumption could plausibly have notable effects on global markets.

This conjecture faces several difficult problems. First, the idea of an "Asian income shock" in 2007–08, when prices first jumped, is implausible on its face. By then, news of fast growth prospects in India and China was hardly shocking; growth had been historically rapid and sustained for years. Second, Indian consumption of cereals actually appears to have been falling on a per capita basis during this period of sustained, unprecedented average increases in Indian income (Deaton and Drèze 2009). This observation is itself difficult to understand, but there is little doubt that an Indian grain demand surge did not happen. In China, the evidence is difficult to validate, but grain and meat demands generally appear to have risen remarkably little recently, on a per capita basis, given the high rates of income increase. Third, as other observers have noted, India and China have had scant engagement in the global grain markets in recent years (for example, Abbott, Hurt, and Tyler 2008; Baffes and Haniotis 2010; Heady and Fan 2008). Indeed, based on data from the Production, Supply, and Demand (PSD) Online website at the US Department of Agriculture (http://www.fas.usda.gov/psdonline/), the net exports from India and China of calories in the three major grains have been positive but never much more than 2 percent of world consumption since 1996, except in 2002–2003 when a jump in corn exports from China actually helped *prevent* a price spike in that period, a positive role that has received scant credit in discussions of the global grains market. To put it another way, supply of grains in China and India has generally been expanding with demand, and China and India have not greatly altered the role they play in international grains trade.

There is one important exception to the above argument. China has been increasing its imports of soybeans rapidly, and the calorie value of those imports is

[9] Such anticipation effects are discussed in Carter, Rausser, and Smith (2012).

Figure 8
Percentage Production Change from Prior Year: World less China



Note: World and China production data for marketing years 1960/61 to 2013/14 are from the Production, Supply, and Demand (PSD) Online database of the US Department of Agriculture.

significant relative to global grains trade. (Exactly what China is doing with those soybeans is not obvious; there are reports that stocks of domestic-origin soy production in China are rising.) On the other hand, this extra soybean demand is balanced by a surge in world supply of palm oil, a substitute for soybeans and canola oil in cooking and biodiesel, that is somewhat larger than the increase in Chinese imports in caloric terms. So despite unsolved puzzles regarding declining Indian per capita cereal consumption and increasing Chinese soy imports, an Asian income shock does not seem to be a major cause of recent grain price increases, although strong competition for land from oilseed producers no doubt limited acreage available for expansion of grain production.

Grain Harvest Shortfalls

An unusually prolonged drought in Australia, drought and fires in Russia and other production shocks figure prominently in discussions of the price spikes in 2007–08. Of course, at any given time, it is usual to observe production problems somewhere in global agriculture. Have these problems been unusually bad in the years since 2005, and are shortfalls increasing in severity or persistence, as many public discussions of this topic seem to imply? Figure 8, showing year-on-year changes in production of each grain, presents the answer quite clearly.

There had been no really significant global production downturns since 1997 until the bad corn harvest of the 2012 crop year. The most volatile harvests were in corn between the years 1975 and 1996. The data suggest that production of major crops is indeed becoming less variable relative to its mean as yields increase, contrary to frequent opposing claims.[10] There is no hint of increased downturns due to global warming in the data. Conjectures that higher carbon dioxide levels have actually increased recent production might be more plausible, but they give no help to those arguing that low harvests explain recent grain price rises.

Prices of Fertilizer, Energy, and Other Inputs

The prices of inputs, in particular fertilizers and fuels, are another popular choice as a driver of market volatility. However the facts say otherwise, at least for the United States. The ratio of a US Department of Agriculture index of prices received by US farmers to an index of prices they paid for inputs, interest, wages, and taxes, in real dollars, has been higher than in 2005 for every year through 2012. Indeed, this change is reflected in the soaring cash rents and land values in the Midwest corn belt (for discussion, see Zulauf and Rettig 2013a, b). Any increases in agricultural input prices have been considerably outstripped by the surge in grain prices.

Interest Rates

Before the grain price surge began, Frankel (2006) pointed out that interest rates are a potentially important influence on commodity prices. In the context of the storage model, an interest rate shift can cause the storage demand function to shift and price to jump. During the financial crisis starting in mid-2008, credit rationing or a jump in interest rates available to traders might well have influenced the drop in commodity prices at that time, by reducing storage and increasing consumption. Since then interest rates have remained very low, and so are not a major explanation of recent grain market behavior.

Bubbles in Prices?

A last rationale for the apparently changed behavior of grain markets in recent years is a story of price bubbles and market manipulation. The story is that powerful investors have chosen to raise the stocks of grains in an attempt to create a shortage and drive up prices, so that they can later sell the stocks at a higher price.

Many discussions of the grain price spikes since 2006 identify "bubbles" in price behavior. There is a common opinion that bubbles are easy to recognize after they have occurred, and that they represent irrational behavior of a kind that might be discouraged or mitigated by appropriate policies against speculation. There is a prevalent opinion that the price spikes in grains in 2007–2008 were speculative bubbles induced by financial flows into grain markets. Timmer (2010, p. 3) states the thesis nicely:

> The actual price panic that resulted, however, had little rationale in the fundamentals of supply and demand. Speculative fervor spread from the crude oil

[10] For example, the coefficient of variation of yield from a linear trend between 1996 and 2011 was at least 29 percent lower than between 1940 and 1995 for wheat, rice, soy, and corn in the United States, and the decrease was significant at the 5 percent level for corn and rice (Zulauf and Hertzog 2011).

> and metals markets to agricultural commodity markets . . . . Prices spiked, first for wheat, then for corn. And then they collapsed when the speculative bubble burst. . . . There is a clear case to be made that the sudden spike in wheat and corn prices was heavily influenced by financial speculation.

Similarly, Piesse and Thirtle (2009) infer bubbles in the 2007–2008 price behavior of soybeans, wheat, and corn and a "panic" in the case of rice, while Headey and Fan (2010, p. xiii) single out the rice price surge as "almost entirely a bubble phenomenon." One source of confusion is that economists find bubbles difficult to define. Brunnermeier (2008) includes a key feature of most definitions offered by finance economists: "Bubbles refer to asset prices that exceed an asset's fundamental value because current owners believe they can resell the asset at an even higher price." But price behavior consistent with this definition can be stationary (Bobenrieth, Bobenrieth, and Wright 2002) and can exhibit behavior typical of prices of grains with occasional spikes but no corresponding troughs (Bobenrieth, Bobenrieth, and Wright 2008). There may be frequent "runs" in which price rises faster than the interest rate (the criterion for "exuberance" in Phillips, Wu, and Yu 2011), then collapse. Indeed, it is not possible to establish the existence of a bubble by rejecting the "no bubble" null hypothesis (Bobenrieth, Bobenrieth, and Wright forthcoming).

Examples of Losers and Winners: Global Poor Nonfarmers and US Landholders

Biofuels policy, including the mandatory diversion of grains and oilseeds from food and feed to transport fuel, led by the United States and the European Union, and followed by many other countries rich and poor, has in effect replicated a classic policy in the agricultural sector of transferring wealth from consumers to producers. Output of calories is transferred from a market with low demand elasticity—the global market for human food and animal feed—to another market with a very high price elasticity—in this case, the global market for motor vehicle fuels. The result of this shift is that price rises greatly in the first market, price falls very little in the second, and producer revenues increase.

I have found nothing in the literature on how this worked out in practice for consumers beyond simulated changes in the number of people below certain international poverty measures. Despite the concerted efforts of countries, notably India, and international organizations including the UN Food and Agriculture Organization (FAO) and the World Bank, we apparently actually know embarrassingly little about how food consumption and prices have evolved at the individual or household level on a worldwide basis. We do know that prices vary widely between countries, and that many grain importing and exporting nations buffered the initial shocks of higher food prices in 2007–08 at substantial budgetary cost, aggravating the price fluctuations in the global market (Anderson, Ivanic, and Martin 2013).

However, for African economies including Kenya, Ethiopia, and Senegal, recent evidence indicates that corn and wheat prices seem to have risen about as much as or more than global prices between 2005 and 2011. Meanwhile, in China, rice prices appear to have gradually risen to above the new higher global price levels, and, in India, close to global levels (Baltzer 2013).

Overall, it seems reasonable to conclude that there has indeed been something like a doubling of the real price paid by the world's landless poor for the world's dominant calorie staples since 2004—along with similar induced increases in prices of many other food mainstays. This change is barely perceptible to the citizens of wealthy countries, who spend a much lower share of their income on calories. Farmers lose if they are net grain consumers, as most poorer farmers are. Landless laborers lose unless their wages rise enough to pay for the more expensive grain. The urban poor unambiguously lose from higher grain prices.

If the absolute number of urban poor living below the $1 per day poverty level have not changed since 2002, then using Chen and Ravallion (2007, table 5, p. 16761) we can estimate their number at 282 million—a conservative estimate given ongoing urbanization of global poverty. Rough back-of-the-envelope calculations suggest that this group (not much smaller than the population of the United States) lost at least $5 billion in 2012 from cereal price rises from the levels prevailing in 2004 before the expansion of biofuels.[11] Rural landless groups and small farmers who are net food buyers likely lost much more in aggregate than this much smaller urban group living below the poverty line. However the effects are much more difficult to calculate accurately, because higher food prices might positively affect incomes of workers working on larger farms, thus partly offsetting the loss from higher cost of cereal consumption.

In real 2012 dollars, real increase in value of US farm operations from 2004 to 2012 is at least $800 billion (USDA 2013; Zulauf and Rettig 2013a), far higher than total US official overseas development aid expenditures over this period, which have never been above $40 billion per year. However, this amount is only a small fraction of the aggregate global wealth transfer: for example, it does not count the gains to landholders in other countries, nor to agricultural input providers and ethanol refiners globally. Most of the global transfers from higher grain prices no doubt remain within national borders, from landless consumers and small landholders, to commercial farmers, agricultural input providers, and others controlling key assets in the marketing chain.

[11] Local price information summarized in Baltzer (2013) can support an increase of around $200 per ton of rice between 2012 for this group and perhaps $180 per ton of wheat (reflecting a lower wheat price rise in India). If the average consumer in this group consumes the cereal calories reported for the urban consumers in a set of available UN Food and Agriculture Organization (FAO) studies, or has the cereal calorie consumption of the very poor in India (Deaton and Drèze 2009, p. 47, table 2), such a consumer paid roughly $22 more for annual cereal consumption in 2012 than in 2004, assuming no reduction in grain calorie consumption, and the aggregate loss to all of this poor urban group is around $6.2 billion. If they were forced by the income effect to cut calorie consumption, the increased expenditure was smaller, but the human costs were very likely even higher.

Some economists warned back in 2007 and 2008 that the biofuels mandates would affect food consumers around the world, but their words went unheeded. For example, Runge and Senauer (2007) made some of the points presented here. They noted that in March 2007, soon after President George W. Bush announced a major expansion of ethanol biofuels, corn futures rose to the highest level in ten years, and also that wheat and rice prices had risen to decade highs because they were substitutes for corn in production and consumption. With impressive foresight, they argued: "By putting pressure on global supplies of edible crops, the surge in ethanol production will translate into higher prices for both processed and staple foods around the world."

Similarly, a World Bank paper by Mitchell (2008) argues that the most important factor in the rapid rise in food prices that had been a burden on the poor in developing countries was "the large increase in biofuels production in the U.S. and the EU. . . . Without these increases, global wheat and maize stocks would not have declined appreciably, oilseed prices would not have tripled, and price increases due to other factors, such as droughts, would have been more moderate. Recent export bans and speculative activities would probably not have occurred because they were largely responses to rising prices." Abbott, Hurt, and Tyler (2008) was another informative early paper that in a comprehensive review identified biofuels demand as one of three major factors in grain price increases.[12] The most important analytical element in this paper that I have added to Mitchell (2008) and Abbott, Hurt, and Tyler (2008) is the discussion of the dynamic response of stocks to a pre-announced path of increased biofuel diversion, which shows why stocks might not fall to smooth a sharp rise in price after a surprising sustained shift in demand.

But many other observers in important institutions with an interest in food—the World Bank (Baffes and Haniotis 2010; Baffes and Dennis 2013), the International Food Policy Research Institute (von Braun and Torero 2009), and the US Department of Agriculture (Trostle, Marti, Rosen, and Westcott 2011)—have, like the Farm Foundation, tended to deemphasize or to ignore biofuels, in favor of a long list of other factors potentially affecting grain prices. "Complexity" and "perfect storm" are words seen frequently in their reports. Surprisingly, academic development economists have not paid much attention to the transfers from poor consumers due to higher food prices: a brief review reveals no papers

[12] However, you would never guess at this conclusion of the study if you read only the Preface by the President of the Farm Foundation, which funded the study. He chose to summarize its conclusions thus: "Today's food price levels are the result of complex interactions among multiple factors—including crude oil prices, exchange rates, growing demand for food and slowing growth in agricultural productivity—as well as the agricultural, energy and trade policy choices made by nations of the world. But one simple fact stands out: economic growth and rising human aspirations are putting ever greater pressure on the global resource base." Later, in his preface to Abbott, Hurt, and Tyler (2009) arguing that the current situation was "remarkably different" than at the time of the 2008 report, he proved able to summarize the latter more succinctly: "Released in July 2008, *What's Driving Food Prices?* identified three major drivers of prices—depreciation of the U.S. dollar, changes in production and consumption, and growth in biofuels production."

on the effects of high grain prices or of biofuels on the global poor in the *Journal of Development Economics* or the *Journal of Development Studies* between 2009 and the writing of this paper in 2013.[13]

Conclusion: The Economics, Politics, and Sustainability of the Regime Change

The behavior of the prices of the major grains reflects their substitutability and their storability. Before the introduction of increasing biofuels mandates in 2005, grain price dynamics reflected the long-run role of crop yield increases in outrunning population increase and the short-run effects of transient supply shocks, which are much more likely to cause price spikes when stocks are low relative to anticipated consumption.

Since 2005, rises in grain price levels have been induced by the sustained increases in demand for grain and oilseed calories. Shifts in demand for biofuels where initiated in Europe and the United States by mandates for grain use in biofuels, along with increases in permissible shares of biofuels used in blending with gasoline or diesel that became attractive when petroleum prices were high. The persistence of the shifts in biofuels demand meant that storage could not buffer consumers as they can during a temporary harvest shortfall.

In this new regime, the world grain market will continue to be sensitive to small shocks, and price levels will remain high overall as long as continued shifts in total calories demanded generated by biofuels demand outrun the expansion of supply. The political economy of biofuels expansion reflects the fact that policies originally widely supported as reducing the emission of greenhouse gases have been captured by the beneficiaries of the large induced wealth transfers.

Will producers and agricultural landowners worldwide be able to ensure that shares of biofuels in transport fuels will continue to rise in the medium term, offseting the yield increases that have traditionally benefited global consumers? Environmentalists have grown skeptical of the claimed reductions in greenhouse gas emissions associated with biofuels; indeed, the net effects of biofuels on emissions are now more widely believed to be at best dubious, due to inevitable induced land use changes (Searchinger et al. 2008) that increase greenhouse gas emissions. In the United States, the expansion of biofuels is currently restricted by rules that limit the use of ethanol in regular gasoline (the so-called "blend wall"), in combination with the fall in demand for gasoline due to reduced driving and higher fuel efficiency of automobiles. Further, the Environmental Protection Agency has proposed to modify regulations in a way that reduces expected growth in corn ethanol demand.

[13] Some papers published in agricultural economics journals addressed this topic, including an IFPRI study by Headey and Fan (2008) and World Bank studies such as Ivanic and Martin (2008).

The loss of support from environmentalists, and the proposed EPA ruling to reduce the renewable fuels standard might be signs that the expansion of biofuels will slow or even reverse itself at least until we have second generation biofuels that do not compete with food production, and more effectively reduce greenhouse gas emission. However the biofuels lobbies in Europe and the United States remain strong and influential.[14] Governments in the United States and the European Union have the power to allow expansion of corn ethanol and other biofuels sufficient to outpace any feasible domestic grain supply expansion if petroleum prices remain high. They can, for example, continually increase the proportion of biofuels approved for blending with regular gasoline or diesel fuels far above the current 10 percent in the United States. If they do, many governments in developing countries are likely to follow their lead, as they have in the past. The roughly $800 billion increase in US farm real estate values since the start of the new biofuels regime is evidence that land investors expect governments to continue to allow biofuels demand for grain to expand, regardless of the effects on the environment and on poor grain consumers.

■ *This paper relies on research supported by the Agricultural Marketing Systems (AMIS) Initiative of the GZO, and by the Energy Biosciences Institute, funded by BP, and draws on results produced in collaboration with Eugenio Bobenrieth, Juan Bobenreith, Carlo Cafiero, and Di Zeng.*

References

Abbott, Philip C., Christopher Hurt, and Wallace E. Tyler. 2008. *What's Driving Food Prices?* Farm Foundation Issue Report.

Abbott, Philip C., Christopher Hurt, and Wallace E. Tyler. 2009. *What's Driving Food Prices?* Farm Foundation Issue Report.

Amezaga, J. M., S. L. Boyes, and J. A. Harrison. 2010. "Biofuels Policy in the European Union." 7th International Biofuels Conference, February, New Delhi.

Anderson, Kym, Maros Ivanic, and Will Martin. 2013. "Food Price Spikes, Price Insulation and Poverty." NBER Working Paper 19530, October.

Baffes, John, and Allen Dennis. 2013. "Long Term Drivers of Food Prices." World Bank Development Prospects Group Policy Working Paper 6455, May.

Baffes, John, and Tassos Haniotis. 2010. "Placing the 2006/08 Commodity Price Boom into Perspective." Policy Research Working Paper 5371, World Bank.

Baltzer, Kenneth. 2013. "International to Domestic Price Transmission in Fourteen Developing Countries during the 2007-08 Food Crisis." WIDER Working Paper 2013/031.

Bellman, Richard. 1958. *Dynamic Programming.*

Bessembinder, Hendrik, Jay F. Coughenour, Paul J. Seguin, and Margaret Monroe Smoller. 1995. "Mean Reversion in Equilibrium Asset

[14] Many environmentalists supported recent European Union votes to cap biofuels use in transport fuels at 6 percent or less, and to ensure that indirect land use effects be considered, but I understand that these votes appear likely to be effectively nullified in response to pressure from biofuels beneficiaries. In the United States, the proposed EPA ruling is open for comment and is being vigorously opposed by farm and biofuels lobbies.

Prices: Evidence from the Futures Term Structure." *The Journal of Finance* 50 (1): 361–375.

Biofuels Digest. 2011. "Biofuels Mandates around the World." July 21. http://www.biofuelsdigest.com/bdigest/2011/07/21/biofuels-mandates-around-the-world/.

Bobenrieth H., Eugenio S. A. , Juan R. A. Bobenrieth H, and Brian D. Wright. 2002. "A Commodity Price Process with a Unique Continuous Invariant Distribution Having Infinite Mean." *Econometrica* 70(3): 1213–1219.

Bobenrieth H., Eugenio S. A., Juan R. A. Bobenrieth H., and Brian D. Wright. 2008. "A Foundation for the Solution of Consumption-Saving Behavior with a Borrowing Constraint and Unbounded Marginal Utility." *Journal of Economic Dynamics and Control* 32(3): 695–708.

Bobenrieth H., Eugenio S. A., Juan R. A. Bobenrieth H., and Brian D. Wright. Forthcoming. "Bubble Troubles? Rational Storage, Mean Reversion and Runs in Commodity Prices." Chap. 5 in The Economics of Food Price Volatility, edited by Jean-Paul Chavas, David Hummels, and Brian Wright. University of Chicago for the National Bureau of Economic Research.

BP. 2012. *Statistical Review of World Energy.*

Brown, Lester. 2012. *Full Planet, Empty Plates: The New Geopolitics of Food Scarcity.* New York and London: W.W. Norton.

Brunnermeier, Markus. 2008. "Bubbles." *New Palgrave Dictionary of Economics.* Basingstoke, Hampshire; New York: Palgrave Macmillan.

Cafiero, Carlo, Eugenio S. A. Bobenrieth H., Juan R. A. Bobenrieth H., and Brian D. Wright. 2011. "The Empirical Relevance of the Competitive Storage Model." *Journal of Econometrics* 162(1): 44–54.

Cafiero, Carlo, Eugenio S. A. Bobenrieth H., Juan R. A. Bobenrieth H., and Brian D. Wright. 2013. "Can Simple Storage Arbitrage Explain Commodity Price Dynamics? Evidence from Sugar Prices." Unpublished paper.

Carter, Colin A., Gordon C. Rausser, and Aaron Smith. 2011. "Commodity Booms and Busts." *Annual Review of Resource Economics* 3: 87–118.

Carter, Colin, Gordon Rausser, and Aaron Smith. 2012. "The Effect of the US Ethanol Mandate on Corn Prices." *Department of Agricultural and Resource Economics, UC Davis.* http://agecon.ucdavis.edu/people/faculty/aaronsmith/docs/Carter_Rausser_Smith_Ethanol_Paper_submit.pdf.

Chai, Andreas, and Alessio Moneta. 2010. "Engel Curves." *Journal of Economic Perspectives* 24(1): 225–40.

Chen, Shaohua, and Martin Ravallion. 2007. "Absolute Poverty Measures for the Developing World, 1981–2004." *PNAS* 104(43): 16757–62.

Chambers, Marcus J., and Roy E. Bailey. 1996. "A Theory of Commodity Price Fluctuations." *Journal of Political Economy* 104(5): 924–57.

Deaton, Angus, and Jean Drèze. 2009. "Food and Nutrition in India: Facts and Interpretations." *Economic & Political Weekly* 44(7): 42–65.

Deaton, Angus, and Guy Laroque. 1992. "On the Behaviour of Commodity Prices." *Review of Economic Studies* 59(1): 1–23.

Deaton, Angus, and Guy Laroque. 1995. "Estimating a Nonlinear Rational Expectations Commodity Price Model with Unobservable State Variables." *Journal of Applied Econometrics* 10(S1): S9–S40.

Deaton, Angus, and Guy Laroque. 1996. "Competitive Storage and Commodity Price Dynamics." *Journal of Political Economy* 104(5): 896–923.

Frankel, Jeffrey A. 1986. "Expectations and Commodity Price Dynamics: The Overshooting Model." *American Journal of Agricultural Economics* 68(2): 344–48.

Frankel, Jeffrey A. 2006. "The Effect Of Monetary Policy on Real Commodity Prices." NBER Working Paper 12713, December.

Galtier, Franck. 2013. "Managing Food Price Instability: Critical Assessment of the Dominant Doctrine." *Global Food Security* 2(2): 72–81. http://www.sciencedirect.com/science/article/pii/S2211912413000035.

Gardner, Bruce L. 1979. *Optimal Stockpiling of Grain.* Lexington Massachusetts: Lexington Books.

Gouel, Christoff. 2012. "Agricultural Price Instability: A Survey of Competing Explanations and Remedies." *Journal of Economic Surveys* 26 (1): 129–156.

Gustafson, Robert L. 1958. *Carryover Levels for Grains: A Method for Determining Amounts that Are Optimal under Specified Conditions.* Vol. 1178. US Department of Agriculture.

Headey, Derek, and Shenggen Fan. 2008. "Anatomy of a Crisis: The Causes and Consequences of Surging Food Prices." *Agricultural Economics* 39 (s1): 375–391.

Headey, Derek, and Shenggen Fan. 2010. "Reflections on the Global Food Crisis. How Did It Happen? How Has It Hurt? And How Can We Prevent the Next One?" International Food Policy Research Institute Research Monograph 185.

Ivanic, Maros, and Will Martin. 2008. "Implications of Higher Global Food Prices for Poverty in Low-Income Countries-Super-1." *Agricultural Economics* 39(s1): 405–416.

Krugman, Paul. 2008. "Grains Gone Wild." *New York Times,* April 7.

Mitchell, Donald. 2008. "A Note on Rising Food Prices." Policy Working Paper 4682, Development Prospects Group, World Bank.

Muth, John F. 1961. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29(3): 315–335.

Nerlove, Marc. 1956. "Estimates of the Elasticities of Supply of Selected Agricultural Commodities." *Journal of Farm Economics* 38(2): 492–509.

Phillips, Peter C. B., Yangru Wu, and Jun Yu. 2011. "Explosive Behavior in the 1990s NASDAQ: When Did Exuberance Escalate Asset Values?" *International Economic Review* 52(1): 201–26.

Piesse, Jenifer, and Colin Thirtle. 2009. "Three Bubbles and a Panic: An Explanatory Review of Recent Food Commodity Price Events." *Food Policy* 34(2): 119–29.

Roberts, Michael J., and Wolfram Schlenker. 2009. "World Supply and Demand of Food Commodity Calories." *American Journal of Agricultural Economics* 91(5): 1235–42.

Roberts, Michael J., and Wolfram Schlenker. 2013. "Identifying Supply and Demand Elasticities of Agricultural Commodities: Implications for the US Ethanol Mandate." *American Economic Review* 103(6): 2265–95.

Roberts, Michael J., and A. Nam Tran. 2012. "Commodity Price Adjustment in a Competitive Storage Model with an Application to the US Biofuel Policies." 2012 Annual Meeting, August 12–14, 2012, Seattle, Washington, Agricultural and Applied Economics Association.

Runge, C. Ford, and Benjamin Senauer. 2007. "How Biofuels Could Starve the Poor." *Foreign Affairs*, May/June issue.

Runge, C. Ford, and Benjamin Senauer. 2008. "How Ethanol Fuels the Food Crisis." *Foreign Affairs*, May 28. http://www.foreignaffairs.com/articles/64915/c-ford-runge-and-benjamin-senauer/how-ethanol-fuels-the-food-crisis.

Scheinkman, Jose A., and Jack Schechtman. 1983. "A Simple Competitive Model with Production and Storage." *Review of Economic Studies* 50(3): 427–41.

Schnepf, Randy. 2006. *European Union Biofuels Policy and Agriculture: An Overview.* CRS Report for Congress RS22404, Congressional Research Service. March 16.

Searchinger, Timothy, Ralph Heimlich, R. A. Houghton, Fengxia Dong, Amani Elobeid, Jacinto Fabiosa, Simla Tokgoz, Dermot Hayes, and Tun-Hsiang Yu. 2008. "Use of U.S. Croplands for Biofuels Increases Greenhouse Gases through Emissions from Land-Use Change." *Science* 319(5867): 1238–40.

Stock, James H., and Francesco Trebbi. 2003. "Who Invented Instrumental Variable Regression?" *Journal of Economic Perspectives* 17(3): 177–94.

Swinnen, Johan, and Pasquamaria Sqiucciarini. 2012. "Mixed Messages on Prices and Food Security." *Science* 335(6067): 405–406.

Timmer, C. Peter. 2010. "Reflections on Food Crises Past." *Food Policy* 35(1): 1–11.

Trostle, Ronald, Daniel Marti, Stacey Rosen, and Paul Westcott. 2011. "Why Have Food Commodity Prices Risen Again?" Outlook No. WRS-1103, Economic Research Service, United States Department of Agriculture, June.

US Department of Agriculture (USDA), Interagency Agricultural Projections Committee. 2008. *Agricultural Projections to 2017.* Long-term Projections Report OCE-2008-1. February.

von Braun, Joachim, and Maximo Torero. 2009. "Exploring the Price Spike." *Choices* First Quarter, 24(1).

Williams, Jeffrey C. and Brian D. Wright. 1991. *Storage and Commodity Markets.* Cambridge, UK: Cambridge University Press.

Wright, Brian D., and Jeffrey C. Williams. 1982a. "The Economic Role of Commodity Storage." *Economic Journal* 92(367): 596–614.

Wright, Brian D., and Jeffrey C. Williams. 1982b. "The Roles of Public and Private Storage in Managing Oil Import Disruptions." *Bell Journal of Economics* 13(2): 341–53.

Wright, Brian D., and Jeffrey C. Williams. 1984. "The Welfare Effects of the Introduction of Storage." *Quarterly Journal of Economics* 99(1): 169–92.

Zilberman, David, Gal Hochman, Deepak Rajagopal, Steve Sexton, and Govinda Timilsina. 2013. "The Impact of Biofuels on Commodity Food Prices: Assessment of Findings." *American Journal of Agricultural Economics* 95(2): 275–81.

Zulauf, Carl, and Evan Hertzog. 2011. "Biotechnology and Variation in Average U.S. Yields." *Farmdoc Daily*, December 14. http://farmdocdaily.illinois.edu/2011/12/biotechnology-and-variation-in.html.

Zulauf, Carl, and Nick Rettig. 2013a. "Comparing Current and 1970 Farm Prosperity: Cash Income and Real Estate." *Farmdoc Daily,* April 5. http://farmdocdaily.illinois.edu/2013/04/comparing-current-and-1970-far.html.

Zulauf, Carl, and Nick Rettig. 2013b. "Comparing Current and 1970 Farm Prosperity: Crop Prices." *Farmdoc Daily,* March 22. http://farmdocdaily.illinois.edu/2013/03/current-1970-farm-prosperity-crop-prices.html.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 99–120

Agricultural Biotechnology: The Promise and Prospects of Genetically Modified Crops†

Geoffrey Barrows, Steven Sexton, and David Zilberman

For millennia, humans have modified plant genes in order to develop crops best suited for food, fiber, feed, and energy production. The earliest efforts, far predating Gregor Mendel's 19th-century discoveries on trait inheritance, involved the selective breeding of plants with desirable characteristics, but the recombination of DNA in offspring was random. Consequently, plant breeding often took decades and frequently yielded crop varieties with unforeseen and undesirable properties. Today, conventional plant breeding remains inherently random and slow, constrained by the availability of desirable traits in closely related plant species. In contrast, agricultural biotechnology employs the modern tools of genetic engineering to reduce uncertainty and breeding time and to transfer traits from more distantly related plants.

Arguments in support of and in opposition to the use of genetically engineered seeds have changed little since the technology emerged in the 1980s. On one side, critics express concerns that the technology imposes negative environmental effects and jeopardizes the health of those who consume the "frankenfoods." On the other side, supporters emphasize potential gains from boosting output and lowering food prices for consumers. They argue that such gains are achieved contemporaneous with the adoption of farming practices that lower agrochemical use and lessen soil

■ *Geoffrey Barrows is a Ph.D. student in the Department of Agricultural and Resource Economics, University of California, Berkeley, California. Steven Sexton is Assistant Professor of Agricultural and Resource Economics, North Carolina State University, Raleigh, North Carolina. David Zilberman is Professor of Agricultural and Resource Economics, University of California, Berkeley, California. Their email addresses are gmb103@berkeley.edu, steven.sexton@ncsu.edu, and zilber11@berkeley.edu.*

† To access the disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.1.99 doi=10.1257/jep.28.1.99

erosion. Although the arguments have changed little since the 1980s when the first generation of genetically engineered crops were created to reduce pest damage, genetic plant engineering became the most rapidly adopted agricultural innovation in history, planted to a cumulative 1.25 billion acres from commercialization in 1996 to today. Roughly 10 percent of all cropland employs the technology in spite of lingering public concerns (James 2011).

The extensive experience with agricultural biotechnology since 1996 provides ample evidence with which to test the claims of supporters and opponents and to evaluate the prospects of genetic crop engineering. In this paper, we begin with an overview of the adoption of the first generation of agricultural biotechnology crops. We then look at the evidence on the effects of these crops: on output and prices, on the environment, and on consumer health. We then consider intellectual property issues surrounding this new technology; a common complaint from critics is that much of the basic research supporting genetic plant engineering was conducted by the public sector, yet the technologies that followed were developed and commercialized by private firms that enjoy intellectual property protections, upending a decades-long tradition of public sector seed development.

We argue that while a number of the environmental issues with genetically engineered seeds warrant scrutiny, the accumulated experience with the first wave of agricultural biotechnology has generated considerable benefits to consumers and the environment. While the next wave of genetic engineering has potential to improve crop response to climate change and boost the nutrient density of staple crops, attention must be paid to the unique risks each new trait may pose. Policy must also seek to ensure that innovation is not unnecessarily burdened and that those who stand to benefit most from the technology—the poor in developing countries—are not neglected. Agriculture is challenged at the outset of the 21st century to feed, clothe, and fuel a world population growing in size and wealth. The history of modern farming lends optimism that the challenge can be overcome, but with the sources of historic growth—mechanization, conventional plant breeding, agrochemicals, and irrigation—reaching diminishing returns, a commitment to new technologies like agricultural biotechnology is needed.

Adoption of Insect-Resistant and Herbicide-Tolerant Seed

The first generation of agricultural biotechnology introduced insect-resistant and herbicide-tolerant traits into four principle row crops. The insect-resistant trait, introduced into corn, cotton, and soybeans, caused crop plants to produce the naturally occurring chemical *Bacillus thuringiensis* (Bt), which is toxic to common agricultural pests, such as the European corn borer, but harmless to humans and relatively environmentally benign. In producing the toxin, which has been applied to plants for nearly a century and is employed in modern organic farming, insect-resistant crop plants rebuff pests without farmers' application of chemicals. The herbicide-tolerant crops express tolerance to glyphosates, a class of broad-spectrum, low-toxicity herbicides that include Roundup®, a Monsanto product employed also in residential settings. Such tolerance, introduced into corn, soybeans, and canola, allows farmers to control weeds more easily. In the absence of herbicide-tolerant varieties, farmers must rely more heavily on either controlling weeds before crop emergence: for example, by repeatedly tilling the soil in a process that causes erosion, or applying relatively more toxic "narrow spectrum" chemicals that can target weeds without affecting post-emergent crops.[1]

Genetically engineered crops were quickly adopted following commercialization in 1996. By 2010, genetically engineered crops were annually planted across 140 million hectares in 29 countries. The technology was adopted on 42 percent of land planted to the four principal genetically engineered crops: corn, soybean, cotton, and rapeseed. Twenty percent of all cropland was planted to genetically engineered seed. Genetically engineered seed was planted to 70 percent of total soybean area, 25 percent of total corn area, 60 percent of total cotton area, and 20 percent of total rapeseed area. The majority of genetically engineered crop area was concentrated among a few countries that aggressively adopted the technologies: the United States and Brazil planted 85 percent of genetically engineered corn, and, with Argentina, 92 percent of genetically engineered soybean. Ninety percent of genetically engineered cotton was planted in India, China, and the United States, while Canada alone planted 85 percent of genetically engineered rapeseed. The area planted to each of these genetically engineered crops is reported in Table 1 for the top adopting countries.

Agricultural biotechnology adoption occurs along an intensive margin as conventional seed is replaced by genetically engineered seed of the same crop. Adoption also occurs along an extensive margin, when natural land and land previously planted to other crops is recruited into production of a genetically engineered crop. Analysis of supply and price effects of genetically engineered crop adoption, as well as environmental impacts, depends critically on rates of adoption across the two margins. Supply effects, for instance, will be greater where the introduction of genetically engineered crops induces farming on marginal lands that were previously unfarmed. But such recruitment of unfarmed lands into agricultural production may come at a cost of environmental damage associated with land-use change.

Figure 1 plots world aggregate acreage over time of four crops with genetically engineered varieties, decomposing total crop area into area planted to traditional seed technology, area planted to genetically engineered seed along the intensive margin, and area planted to genetically engineered seed along the extensive margin. In the absence of records documenting the prior use of land planted to agricultural biotechnology, the decomposition of crop area into these component parts follows an algorithm we develop in Barrows, Sexton, and Zilberman (2013). The algorithm employs aggregate, country-level changes in total and genetically

[1] Glyphosate has a US Environmental Protection Agency (EPA) Toxicity Class of III (on a I to IV scale, where IV is least dangerous) for oral and inhalation exposure. EPA requires that products containing glyphosate carry a label that warns against oral intake, mandates the use of protective clothing, and instructs users not to re-enter treated fields for at least four hours.

Table 1
Genetically Engineered Area Harvested in 2010
(millions of hectares)

Cotton		*Soybean*	
India	9.4	United States	29.4
United States	4.1	Brazil	18.4
China	3.5	Argentina	18.0
Argentina	0.6	Paraguay	2.7
Rest of world	1.3	Rest of world	3.1
Corn		*Rapeseed*	
United States	28.2	Canada	6.1
Brazil	7.5	United States	0.5
Argentina	2.8	Australia	0.1
South Africa	1.9	Rest of world	0.0
Rest of world	1.9		

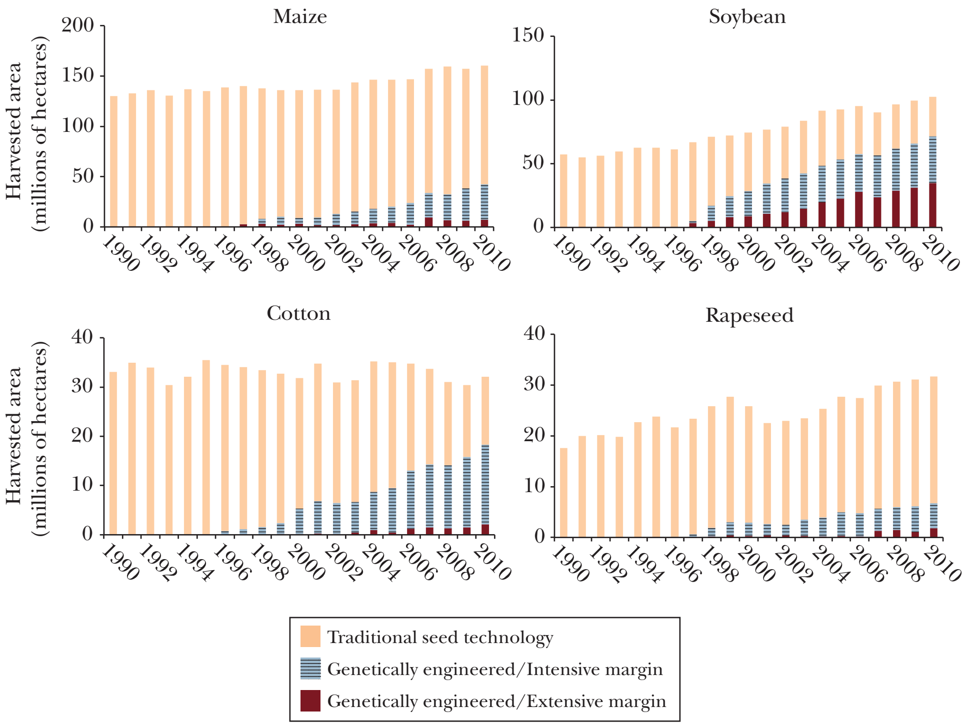
Source: Author's own calculations using data from Graham Brookes.

engineered crop area and a fundamental assumption that if the area planted to a genetically engineered crop is observed to increase from one year to the next, then there were no land transitions out of production of that genetically engineered crop over the same period.

Adoption of genetically engineered cotton, corn, and rapeseed has occurred mostly along the intensive margin, with new seed technology substituting for conventional seed. By contrast, adoption of genetically engineered soybeans has occurred roughly evenly along intensive and extensive margins. Soybean acreage has grown more than 50 percent since the introduction of genetically engineered seed, with most of the gains in Brazil and Argentina.

Agricultural biotechnology adoption along the extensive margin is of considerable interest because of important implications for supply of crops and for environmental quality. Regrettably, there is little information documenting the degree to which production of genetically engineered crops has recruited land from production of other crops or from nonagricultural uses like forest. However, some evidence suggests that growth on the extensive margin also occurs by "double cropping," the practice of planting two crops per growing season instead of just one crop (Trigo and Cap 2003). When double-cropping, farmers produce one early and one late season crop per year, growing in the "shoulder seasons" when pest damage is typically too high for profitable production with conventional seed technology. Seeds engineered with herbicide tolerance are expected to increase double cropping. They permit control of weeds after the crop plant has emerged from the ground, lessening demand for pre-emergence weed control, which typically delays planting long enough to preclude maturation of a follow-on crop. Double-cropping reflects increased annual productivity for a given plot of land, enabling increased

Figure 1
World Area of Four Crops with Genetically Engineered Varieties

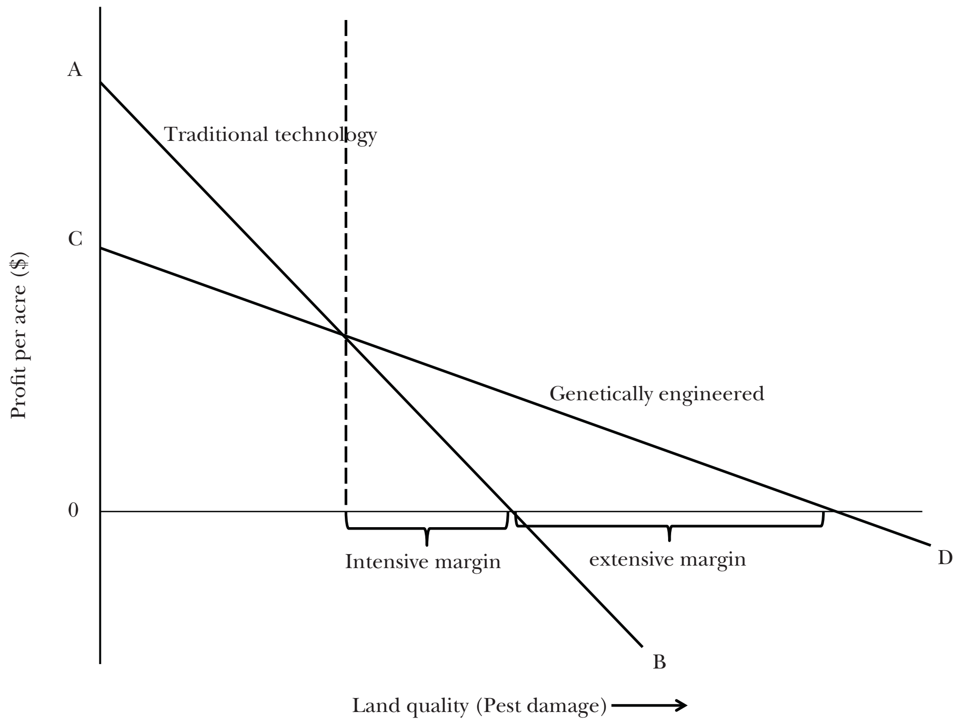


Source: Barrows, Sexton, and Zilberman (2013).
Notes: Figure 1 plots world aggregate acreage over time of four crops with genetically engineered varieties, decomposing total crop area into area planted to traditional seed technology, area planted to genetically engineered seed along the intensive margin, and area planted to genetically engineered seed along the extensive margin.

agricultural output without increasing the footprint of agriculture. There is evidence that adoption of double cropping is correlated with agricultural biotechnology adoption, and that double cropping increased the area planted to soybean in Argentina by 4 million hectares (Trigo and Cap 2003).

To understand agricultural biotechnology adoption patterns and subsequent input and output effects, we model the decision of farmers to use insect-resistant or herbicide-tolerant technology within the damage control framework of Lichtenberg and Zilberman (1986). This framework defines output as the product of potential output and the share of crop not damaged by pests. Pest damage is mitigated by pest control efforts, which may include pesticides applications, agricultural biotechnology adoption, or other agricultural practices. The benefits of adoption are increasing in pest pressure, so, *ceteris paribus*, adoption is more likely in areas characterized by substantial pest problems. When genetically engineered seed replaces chemical applications, the main effect of the adoption is to lower the cost

Figure 2
Adoption of Genetically Engineered Technology



Source: Authors.
Notes: Figure 2 illustrates the damage control model. Land quality is decreasing along the horizontal axis, with low-pest-pressure, high-quality land located on the left and low-quality, high-pest-pressure land on the right. Profit per acre is measured along the vertical axis.

of damage control. When genetically engineered seed are adopted on farms with uncontrolled pest problems, the main effect of adoption is to increase yield (Qaim and Zilberman 2003). In both cases, adoption is more likely in locations with more significant pest problems.

Figure 2 illustrates the damage control model. Land quality is decreasing along the horizontal axis, with low-pest-pressure, high-quality land located on the left and low-quality, high-pest-pressure land on the right. Profit per acre is measured along the vertical axis. Line segment AB depicts profit per acre as a function of land quality (or pest pressure) under the traditional technology, and line segment CD shows profit per acre for genetically engineered technology. In areas characterized by sufficiently low pest pressure, the conventional technology yields higher profits because crop losses avoided by genetically engineered seed are too small to compensate for its higher costs. As the level of pest pressure rises, the gains from adoption of genetically engineered seed increase. For sufficiently high pest pressure, it is only profitable to farm using genetically engineered seed because crop losses from conventional technology are too great. Genetically engineered seed is expected to boost crop production by generating incremental yield gains along the intensive margin and by inducing production along the extensive margin on land that otherwise generated no output.

While profitability has been identified as the major reason for farmer adoption of agricultural biotechnology, the literature has identified other benefits to adoption as well, including reduced exposure to pesticides and reduced effort for pest monitoring that is necessary for optimality of pesticide applications. The magnitude of these benefits is also correlated with pest pressure (Piggott and Marra 2008; National Research Council 2010).

Estimating Effects of Genetically Engineered Seeds on Production and Prices

A number of studies using farm-level data in various countries have found that genetically engineered seeds increase crop yields. The model above suggests that the greatest potential for yield gains from genetically engineered seeds exist in places with high pest pressure and little access to alternative damage control—that is, mainly in low-income developing countries. A corollary implication, discussed in the next section, is that the greatest reductions in insecticide use and toxic herbicide use are expected to occur in developed countries where chemicals were aggressively deployed to reduce damage before the introduction of genetically engineered seed (Qaim and Zilberman 2003). Qaim (2009) summarizes impact studies, finding yield gains of 37, 33, and 24 percent for the insect-resistant genetically engineered Bt cotton in India, Argentina, and China, respectively. Estimated yield gains in the developed world are smaller, with the US yield gains at only 10 percent. A similar pattern holds for insect-resistant Bt corn, with an estimated yield increase of 34 percent in the Philippines, 11 percent in South Africa, 9 percent in Argentina, and only 5 and 6 percent in the United States and Spain, respectively.

Much of the literature has explored the yield benefits of genetically engineered crops in randomized controlled trials settings, where farmer behavior is held constant. The foregoing estimates, therefore, can be considered a pure "gene effect" on yields, reflecting only the damage control benefits afforded by the transgenic trait of the seed. Theory suggests, however, that the observed yield gains should exceed those attributed to the gene effect because diminished crop damage increases the marginal value product of other yield-increasing inputs, like fertilizer, water, labor, and capital. Farmers who adopt genetically engineered crops, are, therefore, likely to also increase use of other inputs that further boost yields.

Of course, the additional crop production due to genetically engineered seed adoption is comprised of yield gains on existing cropland as well as total production on new lands recruited into production by the damage-control savings afforded by the technology. Using the decomposition from Barrows, Sexton, and Zilberman (2013) mentioned in the previous section, aggregate supply and crop price effects are estimated for the four principal crops. Lacking estimates of the causal impact of genetically engineered technology on extensive growth in output, we compute effects under two assumptions providing upper and lower bounds. In the upper bound, we assume all production on the extensive margin is attributed to genetically engineered seed, while in the lower bound, we assume production occurring on the extensive margin would have occurred even without genetically engineered seed. The true effects likely fall somewhere in between, and the range indicates the potential importance of the extensive margin. Across the eight yield estimates for corn appearing in Qaim (2009) and Barrows, Sexton, and Zilberman (2013), the supply effect is estimated to vary from a 2–14 percent increase in total corn supply at the lower bound and a 9–19 percent increase in total corn supply at the upper bound. The smallest estimates are based on estimated yield gains in the United States, Spain, and South Africa. For cotton, lower bound estimates range from 0–25 percent while upper bound estimates range from 5–29 percent. The larger range for cotton reflects the greater variance in estimated yield gains. While there aren't many estimates of yield gains from genetically engineered soybeans, the gains estimated in Sexton and Zilberman (2011) imply a supply effect ranging between 2 and 39 percent, depending on the assumptions about the extensive margin. In Barrows, Sexton, and Zilberman (2013), we offer details on these calculations of extensive growth in output (which should be considered in addition to the intensive estimates above).

Following a standard approach in the literature (for example, de Gorter and Zilberman 1990), the effects of genetically engineered seed on production can be translated into price effects given assumptions about price elasticities of supply and demand. Assuming an own-price demand elasticity of –0.5 and an own-price supply elasticity of 0.3, and attributing equal weight to each of the eight studies reviewed in Barrows, Sexton, and Zilberman (2013), we find that adoption of genetically engineered corn in 2010 lowered prices 13 percent. The expected price decline for cotton was 18 percent. For soybeans, the estimated price decline implied by calculations in Barrows, Sexton, and Zilberman (2013) ranges from 2 to 65 percent.

Crop consumers benefit from lower prices. But all else equal, lower prices hurt farmers. The impact of agricultural biotechnology adoption on farm income has been a subject of considerable concern. A substantial literature has assessed the distribution of benefits and costs from adoption of genetically engineered varieties, mostly using partial equilibrium models (Lapan and Moschini 2004). For example, a National Research Council (2010) survey showed that the relative distribution of benefits among groups varies across products and locations. The share of overall gains accruing to farmers is estimated to be between 5–40 percent. Seed developers capture between 10 to 70 percent of the benefits, and the share of benefits flowing to US consumers is estimated to be between 6 and 60 percent. Consumers in the rest of the world capture 6–30 percent of total benefit. In most cases, seed developers capture less than half of the benefits, with the majority of surplus accruing to farmers and consumers. The differences in estimated outcomes reflect variations in the impacts of different traits at different locations, market structure, and demand and supply parameters.

Genetically engineered seed is relatively easy to adopt as it only requires the substitution of one seed for another. While richer farmers may have been early adopters of the technology (Crost, Shankar, Bennett, and Morse 2007), near 100 percent adoption of Bt cotton in many regions of India and high rates of adoption in Burkina Faso and South Africa yielded gains to smallholders. The contributions of genetic modification technology to food security in a global context are presented in Godfray et al. (2010). A case study by Kathage and Qaim (2012) in India showed that adoption of insect-resistant Bt cotton resulted in a 50 percent increase in profit per hectare and an 18 percent increase in expenditures. Gouse, Pray, and Schimmelpfennig (2004) document gains to smallholders in South Africa from Bt cotton. The total benefits from genetically engineered varieties have been substantial in absolute terms. The global net benefit to producers over the period 1996–2009 was estimated by Brookes and Barfoot (2012) to be $65 billion, of which $30 billion accrued to US producers. Estimation of the overall benefits of genetically engineered technology and the distribution of those benefits is an ongoing area of research.

Environmental Benefits and Risks

Since its inception, genetic plant modification has engendered concerns about adverse impacts on ecosystems and environmental quality. These concerns persist, and include the risks of gene flow to noncrop plants and natural lands, agricultural biodiversity loss, and pesticide resistance build-up for common pesticides. We consider each of these in turn.

Genetic engineering introduces plants with novel phenotypes into existing ecological networks. If these traits escape the farm and spread elsewhere, the effects on surrounding ecosystems could be novel and complex (Wolfenbarger and Phifer 2000). For example, a gene that makes a certain crop heartier might spread to a related weed species and make that weed more invasive. Organic farmers particularly worry about the spread of genetically engineered material from adjacent farms to their fields. Such accidental transfer of transgenic material jeopardizes access to organic markets, which is often premised on sufficient purity. The coexistence of transgenic crops and organic farming typically relies upon the imposition of buffers between crops, the optimal size of which depends upon the risks of material transfer and the costs of impurity (Beckmann, Soregaroli, and Wesseler 2006).

While natural hybridization is common among plants that have close genetic relationships, the ecological effects of gene flows are not well understood. Existing genetically engineered crops are not typically planted in proximity to native relatives, which reduces the risk of traits jumping into wild species. Few crops produced in the United States originated there and have proximal wild relatives. There are no weedy relatives of corn and soybean, the dominant genetically engineered crops in the United States. In areas where wild cotton populations exist in the United States, production of genetically engineered cotton is forbidden to avoid risk from gene flow (Warwick, Beckie, and Hall 2009). Rapeseed has been designated a moderate-risk crop because herbicide-tolerant traits are reported to have spread to wild relatives (Stewart, Halfhill, and Warwick 2003).

Opposition to genetically engineered seed also centers on the technology's propensity to increase monoculture agriculture by reducing its agronomic costs (Pollan 2001). Though monocropping can deplete soil quality and increase pest problems, it affords efficiencies and productivity gains that explain why it is favored by many farmers. The phenomenon began well before the commercialization of genetically engineered seed.

Concerns about resistance build-up are not unfounded. Like other pest control methods, transgenic seed are not immune to evolutionary forces that can induce resistance absent proper management. The National Research Council (2010) reported that resistance to toxins in insect-resistant crops had evolved among only three pest species in the first 14 years of commercial insect-resistant cropping. Bennett, Phipps, Strange, and Grey (2013) report on newer cases of resistance development and their implications. Growing pest resistance to insect-resistant corn in Puerto Rico resulted in the voluntary withdrawal of the genetically engineered seed in 2006 (Tabashnik, Van Rensburg, and Carrière 2009).

At least 10 species of weeds have evolved resistance to glyphosate in herbicide-tolerant fields in the United States due to the nearly exclusive reliance on glyphosate for weed control (Duke and Powles 2009). A growing number of weeds are evolving resistance to glyphosates, but the number of locations in which resistance build-up is problematic is growing faster because of the widespread adoption of herbicide-tolerant crops. Because farmers typically respond to the diminished efficacy of resistant glyphosates by increasing dosage and application frequency and by supplementing with other chemical applications, resistance build-up can adversely impact the natural ecosystem to the extent that farm chemicals drift (National Research Council 2010; Mueller, Mitchell, Young, and Culpepper 2005). Amid resistance build-up, farmers are expected to rely more heavily on tilling operations in lieu of glyphosate applications, potentially worsening soil erosion and water quality and impeding soil carbon sequestration (Mueller et al. 2005).

While the risk of resistance build-up is not unique to genetically engineered seed, the risk is likely greater in transgenic crops than conventional crops. The selection pressure is omnipresent in insect-resistant traits, whereas on traditional crops, the selection pressure can be managed by controlling insecticide applications and varying damage control agents. The efficacy of glyphosates for weed control also suggests that resistance build-up will be greater among herbicide-tolerant crops to which glyphosates can be applied post-emergence.

Pest susceptibility is a common pool resource likely to be underprovided relative to social optimality; no individual farmer faces the full cost of his use of damage-control agents nor accrues the full benefit from his effort to minimize selection pressure. Consequently, regulators mandate that farmers who plant insect-resistant seed also plant refuges of non-insect-resistant crops. Susceptible pests can survive and interbreed with resistant pests in the refuges, and thus maintain the stock of susceptibility. In the United States, such refuges are mandated to be equal in size to at least 20 percent of the area devoted to insect-resistant crop production (Bourguet, Desquilbet, and Lemarié 2005). In some cases, the non-insect-resistant seed are intermingled with the transgenic seed to produce "refuge in a bag."

Given their monopoly status afforded by intellectual property rights, seed companies have incentive to manage resistance in order to preserve the efficacy of their seed technologies. Such an incentive would not exist among competitive suppliers. Seed companies responded to resistance concerns by "stacking" multiple traits in insect-resistant seed. Each trait is designed to target pests differently, lessening the selection pressure. The advent of stacked traits has slowed resistance build-up among insect-resistant traits, but a similar solution is lacking for herbicide-tolerant traits. Seed companies are developing traits that express tolerance to other herbicides, though the other herbicides are less benign than glyphosates.

In spite of environmental risks posed by agricultural biotechnology, theory and empirical evidence suggest genetically engineered crops deliver environmental benefits by saving land and agrochemicals and by maintaining rather than diminishing agricultural biodiversity. While critics assert that agricultural biotechnology has increased pressure to monoculture, genetic engineering can reduce that pressure and maintain crop diversity. It inserts traits into existing crop varieties, modifying them slightly. If the costs of inserting transgenic traits into seed, inclusive of regulatory costs, are low, then the adoption of genetically engineered traits does not necessarily reduce agricultural biodiversity. In fact, much of soybean biodiversity has been sustained (Zilberman, Ameden, and Qaim 2007). Moreover, genetic engineering allows the reintroduction of seed varieties that had been abandoned because of pest damage. Genetic engineering also permits production of differentiated varieties with unique properties, enhancing biodiversity. For example, a new soybean variety with reduced artery-clogging transfats was announced in 2013 (Pollack 2013), and in Bennett, Chi-Ham, Barrows, Sexton, and Zilberman (2013), we enumerate new varieties with desirable health or agronomic properties.

Because insect-resistant seeds substitute for insecticide applications, they are expected to reduce agrochemical applications and limit environmental damage from chemical runoff and drift. Many of the randomized trials surveyed in Qaim (2009) estimated changes in pesticide use in addition to changes in yields. Bt cotton adoption is estimated to reduce pesticide use by 65 percent in China, 47 percent in Argentina, 36 percent in the United States, and 33 percent in South Africa. Cotton is the most pesticide-intensive crop worldwide, and an estimated 128 million kilograms of pesticide applications were avoided worldwide from 1996 to 2007 because of insect-resistant Bt cotton adoption. This is estimated to have reduced the environmental impact of cotton pesticides by 25 percent (Brookes and Barfoot 2006). The reduction in pesticide on genetically engineered corn crops tends to be smaller. While insecticide use in Spain declined 65 percent on insect-resistant Bt corn fields, the United States, South Africa, and the Philippines experienced reductions of only 8, 10, and 5 percent, respectively (Qaim 2009).

In contrast to insect-resistant seeds, which act as substitutes for insecticides, herbicide-tolerant seeds are complimentary with herbicides. However, herbicide-tolerant seed varieties permit the substitution of relatively environmentally benign alternatives for the toxic, narrow-spectrum chemicals used on conventional crops. Thus, while overall quantity of pesticide use may increase with the adoption of herbicide-tolerant crops, the total toxicity of applied chemicals does not (National Research Council 2010). Glyphosates, such as Roundup®, for instance, are less prone than alternative herbicides to leaching, more biodegradable, and less toxic to a variety of animals, including mammals, birds, and fish (Fernandez-Cornejo and McBride 2002; Cerdeira and Duke 2006; Malik, Barry, and Kishore 1989).[2] From 1996 to 2008, the per-acre quantity of glyphosate use in the United States increased roughly proportional to the nearly fivefold increase in herbicide-tolerant soybean acreage (National Research Council 2010). Applications of substitute herbicides, which tended to be more toxic, decreased nearly commensurately. Overall, the quantity of active ingredients applied to cotton and soybean crops increased slightly in the United States since the introduction of genetically engineered seed technology.

While chemical use is shown to increase with adoption of herbicide-tolerant seed, such adoption is also associated with adoption of reduced tillage operations. Tilling is a mechanical form of weed control that causes soil erosion, releases carbon from the soil (which contributes to climate change concerns), and increases farm runoff, which is responsible for nitrification and "dead zones" in the Gulf of Mexico (National Research Council 2010; Brookes and Barfoot 2010). With the adoption of herbicide-tolerant crops, post-emergence glyphosate applications substitute for pre-emergence tilling operations, so that no-till and reduced-till practices increase. In fact, the no-tillage soybean area in the United States doubled from 1996 to 2008 and increased fivefold in Argentina. The no-tillage canola area in Canada tripled from 1996 to 2005, while the no-tillage cotton area increased fivefold in the United States.

Diminished reliance on chemical applications and tillage operations also lowers demand for fuel for farm machinery, which generates cost savings to the farmer and lowers air pollution and greenhouse-gas emissions. By lowering the optimal level of insecticide applications, insect-resistant crops reduce the number of passes farm equipment must make through fields. Though there is no direct evidence of the magnitude of fuel savings associated with adoption of insect-resistant crops, a reasonable approximation is that insecticide use accounts for half of total fuel use on a field (Mitchell, Munk, Prys, Klonsky, Wroble, and De Moura 2006). Assuming a monotonic correspondence between insecticide savings and savings on fuel for insecticide applications, the adoption of insect-resistant seeds generated an estimated 32 percent overall fuel savings on cotton fields in China or corn fields in Spain, and 18 and 4 percent savings on insect-resistant cotton and corn fields in the United

[2] Shaner (2000) compares the active ingredients of glyphosates to other groups of herbicides and documents the relative advantage of glyphosate use.

States, respectively. The yield gains on the intensive margin also avert fuel consumption that would accompany cropland expansion. The use of no-tillage operations on herbicide-tolerant fields lowers fuel use 30–73 percent according to various estimates (Mitchell et al. 2006; Sanders 2000; USDA-NRCS 2008; Jasa 2000).

At the most basic level, the yield gains from genetically engineered seeds reduce the size of the agricultural footprint necessary to produce a given quantity of food, fiber, feed, and fuel. The simulation framework of Sexton and Zilberman (2011) estimates that 20 million additional hectares of land would have been necessary to produce the 2008 harvest of soybeans and corn absent the yield gains from genetically engineered seeds. Because land-use changes account for a leading share of anthropogenic greenhouse gas emissions, this reduction in agricultural land demand represents a meaningful contribution to the abatement of greenhouse gas emissions and also averts biodiversity loss associated with recruitment of native lands into agricultural production.

Human Health Impacts of Genetically Engineered Crops

Consumers have exhibited a willingness to pay to avoid perceived health risks associated with genetically engineered foods (for a survey, see Lusk, Jamal, Kurlander, Roucan, and Taulman 2005) in spite of being generally uninformed about the risks and the benefits of agricultural biotechnology (European Commission Eurobarometer 2000; Pew Charitable Trusts 2001). Moreover, there exists a gap between public perception and scientific knowledge (for example, McHughen 2007; Hoban 1998; Marchant 2001; Marris 2001). Consumer willingness to pay, however, changes in response to new information (Kiesel, McCluskey, and Villas-Boas 2011), and products containing genetically engineered material that delivers health benefits may fetch a premium at market (Colson, Huffman, and Rousu 2011).

One health concern surrounding consumption of genetically engineered food is the possibility that genes inserted into the DNA of plants will be toxic to humans—for example, it may induce allergic reactions. Regulators have sought to prevent the intentional or accidental transfer of genes encoding major allergens into food crops in which they were previously absent (Goodman et al. 2008). There is no evidence that a transgenic gene has introduced allergenicity into a crop or caused the endogenous allergenicity of the crop to increase (Taylor 2006).

A related worry concerns genetic modification techniques that also (inadvertently) select for antibiotic-resistant genes. In particular, there is concern that the efficacy of therapeutic antibiotics may be diminished by consumption of genetically engineered foods derived from crops that contain antibiotic-resistant genes. Some also fear that antibiotic resistance may spread from the crop plant to intestinal or soil microorganisms. Bennett, Phipps, Strange, and Grey (2004) and the European Food Safety Authority (2004), among others, have determined that the risk of horizontal gene transfer from the plant to microorganisms is extremely low and that consequences would be minimal even if the transfers were to occur. European regulators concluded that the antibiotic selectable markers in commercial use present no inherent risk to human health. And new techniques are now available that avoid the problem (Hare and Chua 2002).

The safety paradigm governing regulation of genetically engineered crops is one of "substantial equivalence" that compares novel plants to already existing plants, such as conventionally bred plants for which there is a substantial record of safety. Any characteristic of a new plant that deviates substantially from that of the conventional counterpart is investigated for allergenicity, toxicity, or other unintended effects. The substantial equivalence paradigm is endorsed by the Codex Alimentarius Commission (2003a, 2003b)—a joint Food and Agriculture Organization/World Health Organization food standards program. It is also consistent with a consensus in the scientific community that the recombinant DNA process is not inherently less safe than conventional forms of plant breeding and that the content of crop plants and foods should drive their regulatory scrutiny, not the process by which they were bred (National Research Council 1989, 1994, 1996, 2000, and 2010; European Commission Eurobarometer 2010).

Paarlberg (2010) surveys evidence from the British Medical Association, French Academies of Science, Organisation for Economic Co-operation and Development, and the UN Food and Agriculture Organization in asserting that "GMO [genetically modified organism] foods and crops currently on the market have brought no documented new risks either to human health or to the environment." Yet future traits may pose more risk, so that continued scrutiny of new transgenic crop introductions is warranted. Whereas food safety organizations find nothing inherently unsafe in the process of genetic engineering, they note that new plant characteristics developed in the next wave of agricultural biotechnology could present new risks. At the same time, future generations of biotechnology may well provide food products with higher vitamin or protein content and lower allergenicity, yielding benefits to food consumers.

Even first-generation agricultural biotechnology likely delivers some health benefits that weigh against health risks. Food consumers benefit from reduced exposure to mycotoxins, the toxic and carcinogenic chemicals produced by fungi that colonize crops. By reducing insect pest pressure, insect-resistant crops can reduce fungal contamination. Insect-resistant Bt corn, in particular, has significantly lowered levels of mycotoxins in field trials. The risk from mycotoxin accumulation in food is more severe in developing countries because of ineffective pest control and poor food-storage conditions. Wu (2006) estimated that the economic benefits of reduced contamination due to Bt corn could reach $100 million annually.

Farm workers stand to benefit from reduced exposure to chemicals. Huang, Hu, Rozelle, and Pray (2005) found that fields in China planted with insect-resistant rice reduced pesticide use by 80 percent, which in turn eliminated pesticide-induced illness that afflicted from 3 to 8.3 percent of farmers of conventional rice. Similar health benefits were observed among insect-resistant cotton farmers in China (Huang, Pray, and Rozelle 2002). Though the benefit of reduced pesticide exposure has not been widely investigated, the reduced chemical use associated particularly with insect-resistant seed has been documented widely (see survey by Qaim 2009). The health gains from avoided pesticide applications are likely larger among adopters in developing countries who are less likely to have access to protective equipment.

Finally, the reduction in pesticide use on genetically engineered crops lowers pollution emissions from the production and transportation of agrochemicals and field operations. Lower emissions, in turn, reduce smog formation, toxic particulate matter concentrations, ecotoxicity of water, and water and soil acidification and nitrification (Bennett et al. 2004). Overall, genetically engineered crops are about one-tenth as toxic to the environment as the conventional crop (Bennett et al. 2004).

Intellectual Property Rights and Regulation

Next-generation traits in the biotechnology research pipeline include those that would enhance drought tolerance, nutritional content, shelf life of produce, nitrogen fixation, and adaptability to climate change (Graff, Zilberman, and Bennett 2009; Ronald 2011). The success of future genetically engineered plant technologies, however, depends on innovators' incentives, which are affected by intellectual property rights regimes and safety regulations.

For much of the 20th century, seeds in the United States were often provided by the public sector and distributed to farmers at low cost. In contrast, genetically engineered seed technologies are usually protected by intellectual property law and sold to farmers at monopolistic prices. As some of the relevant patents were developed by university research and subsequently licensed to the private sector, many oppose genetically engineered seeds on the grounds that seed companies are profiting from publicly funded research. Indeed, seed companies (like medical companies) pay to use the rights of patents that were developed by the public sector, but they then spend much larger amounts on development and registration activities (Bennett et al. 2013). Furthermore, the private seed industry was remade, decades before the advent of genetic plant engineering, when hybrid seed varieties were introduced by crossing two parent varieties exhibiting strong phenotypes. Hybrid corn varieties consistently outperformed conventional seed in the 1930s and provided inherent protection of private investment in research and development as the hybrid seed does not express parent traits and, therefore, cannot be saved by farmers (Duvick, Smith, and Cooper 2004).

Growing private sector investment in seed technology, motivated by the potential royalties from innovation, has occurred alongside a prolonged decline in public sector agricultural research and development (Alston, Beddow, and Pardey 2009). Basic research has largely been carried out at universities, which license technologies to seed companies, which conduct comprehensive testing and undertake development and marketing efforts—a division of labor that mirrors what occurs in biomedical research and development. Without investments by major agrochemical companies, such as Monsanto, many genetically engineered crop technologies would not have been developed (Graff, Cullen, Bradford, Zilberman, and Bennett 2003).[3]

Reliance upon private sector intellectual property incentives for agricultural innovations creates some problems. Importantly, huge social welfare gains could be obtained by employing genetic plant engineering to address some of the agronomic problems in developing countries. But seed companies have little incentive to develop traits for these applications because of limited purchasing power. While intellectual property rights have induced private sector research, they have limited the application of these technologies in poor parts of the world. Policies to make patented technologies more widely available in poor countries would enhance both the social benefits of the technology and its perception among the public. Likewise, specialty crop applications are often neglected because the markets are small.

The introduction of intellectual property rights for genomic information has fragmented ownership of traits and enabling technologies across many parties, creating an "anti-commons." Introducing a single new genetically engineered crop can require innovations protected by 40 or more individual patents and license agreements. Public efforts have sought to improve access to intellectual property rights and, in particular, to ensure that developers of philanthropic applications have "freedom to operate" (Atkinson et al. 2003; Koo, Nottenburg, and Pardey 2004). The development of a clearinghouse for agricultural biotechnology property rights serves to improve coordination and access for developing countries and specialty crops (Graff and Zilberman 2001; Graff et al. 2003).

While the patenting of "genes" per se is restricted, there is a strong case for patenting knowledge about functions of genes as it provides a base for useful product innovations that can be developed by the private sector. However, development of such technologies requires access to process innovations. A crucial process innovation in agricultural biotechnology is the use of agrobacterium to "ferry" genes. The rights to this patent are exclusively held by Monsanto. This is in contrast to the rights to use the gene transfer technology in medical biotechnology (the Cohen–Boyer patent), which were nonexcludable and, consequently, utilized by many startups. The limited access to the agrobacterium patent has limited development and commercialization of agricultural biotechnology innovations.

Whereas property rights for genomic information create incentives for private agricultural productivity research, safety regulations create disincentives for new trait introductions by imposing costly testing (Potrykus 2010). In the United States, the cost of regulatory approval for a single variety of genetically engineered seed can reach $15 million. It typically takes ten times more money and ten more years

[3] Global public research in agricultural research and development is $33.7 billion as of 2009, and private agricultural research and development is about 60 percent the amount of public agricultural research and development. However, only a small amount of public research in the United States is allocated to plant genetics. Federal expenditures on this category of research are below half a billion dollars, and the total US (federal and state) spending on plant genetics is below $1 billion, and the United States is the biggest spender on this research (Current Research Information System 2012). On the other hand, Monsanto alone spends about $1.5 billion on research and development annually.

to bring a genetically engineered crop to market than a nongenetically engineered crop (Just, Alston, and Zilberman 2006; Potrykus 2010).

In most of Europe and in many other countries, the production of genetically engineered crops is essentially banned. The perception of many Europeans is that agricultural biotechnology does not benefit them directly but instead only poses health and environmental risks (Paarlberg 2008; Graff, Hochman, and Zilberman 2009). European biotechnology regulations frequently rely on the precautionary principle, which neglects benefit–cost analysis and instead holds that potential risks should be avoided even absent scientific evidence of harm (Foster, Vecchia, and Repacholi 2000). Onerous regulation and de facto bans on genetically engineered crops diminish the market for the technology, lowering the expected return to seed company research and development. For instance, Graff, Zilberman, and Bennett (2009) show that Europe's introduction of a de facto ban on genetically engineered crops in 1999 was associated with a significant contraction in agricultural biotechnology research and development. Investment slowed and the number of field trials fell. As a consequence, hundreds of new traits were stranded in the research and development pipeline, including those that proposed to boost nutrient content of staple crops; extend the shelf lives of food products; enhance the efficiency of animal feed; and protect crops from drought, flood, and saline soils.

The regulatory treatment of agricultural biotechnology deserves careful scrutiny. While existing practices and safety regulations have avoided any major human or environmental health impacts, genetic engineering technologies remain confined to only a handful of commercial crops expressing only a few traits. In the future, the regulatory balance should more evenly balance pre-market testing and post-market review, depending on the novelty of the crop and the nature of the safety risk. For instance, the introduction of an already commercialized trait into a new crop should probably require a less-strenuous review than the introduction of an entirely new trait.

Conclusion

More than a decade and a half since the commercialization of first-generation agricultural biotechnology, concerns about transgenic crop impacts on human and environmental health remain, even though the experience across a cumulative 1.25 billion hectares suggests the relative safety of first-generation genetically engineered seed. The risks posed by agricultural biotechnology warrant continued attention, and new transgenic crops may pose different and bigger risks. Weighing against uncertain risks are benefits from increased food production, reduced insecticide use, and avoided health risks to food consumers and farm workers. At the same time, adoption is shown to increase herbicide use while reducing herbicide toxicity, save land by boosting yields while also making previously unfarmed lands profitable. Adoption benefits food consumers and farmers but also enriches seed companies that enjoy property right protections over new seed varieties. The balance of scientific knowledge weighs in favor of continued adoption of genetically engineered seed, which may explain why some longtime critics have reversed course. For example, Lord Melchett, who was the head of Greenpeace, has been advising biotechnology companies on overcoming constraints to the technology (St. Clair and Frank forthcoming). Mark Lynas, a journalist and organizer of the anti–GM (genetic modification) movement, publicly apologized for helping start the movement in his "Lecture to Oxford Farming Conference" (2013).

Agricultural biotechnology remains regulated by regimes developed at the introduction of the technology. Whereas precaution may have been appropriate before the relative magnitudes of risks and benefits could be empirically observed, accumulated knowledge suggests overregulation is inhibiting the introduction of new transgenic varieties. Regulation also discourages developing-country applications, where benefits are likely greatest. In the future, new genetic traits may promise greater benefits while also posing novel risks of greater magnitudes than existing traits. Efficient innovation and technology adoption will require different and, perhaps, more stringent regulation in the future, as well as continued insights from researchers, including economists, in order to assess evolving costs and benefits.

■ *The authors benefited immensely from comments and suggestions made by Timothy Taylor, John List, David Autor, Chang Tai-Hsieh, and Jeanne Reeves. We also wish to thank Eunice Kim and Scott Kaplan for their assistance. Support for the research leading to this paper was provided by Cotton Inc., and David Zilberman is a member of the Giannini Foundation.*

References

Alston, Julian M., Jason M. Beddow, and Philip G. Pardey. 2009. "Agricultural Research, Productivity, and Food Prices in the Long Run." *Science* 325(5945): 1209–10.

Atkinson, Richard C., Roger N. Beachy, Gordon Conway, France A. Cordova, Marye Anne Fox, Karen A. Holbrook, Daniel F. Klessig, Richard L. McCormick, Peter M. McPherson, Hunter R. Rawlings III, Rip Rapson, Larry N. Vanderhoef, John D. Wiley, and Charles E. Young. 2003. "Public Sector Collaboration for Agricultural IP Management." *Science* 301(5630): 174–75.

Barrows, Geoff, Steve Sexton, and David Zilberman. 2013. "The Impact of Agricultural Biotechnology on Supply and Land-Use." CUDARE Working Paper 1133. University of California, Berkeley, Department of Agricultural and Resource Economics.

Beckmann, Volker, Claudio Soregaroli, and Justus Wesseler. 2006. "Coexistence Rules and Regulations in the European Union." *American Journal of Agricultural Economics* 88(5): 1193–99.

Bennett, Richard, Richard Phipps, Alison Strange, and Peter Grey. 2004. "Environmental and Human Health Impacts of Growing Genetically Modified Herbicide-tolerant Sugar Beet: A Life-cycle Assessment." *Plant Biotechnology Journal* 2(4): 273–78.

Bennett, Alan B., Cecilia Chi-Ham, Geoffrey Barrows, Steven Sexton, and David Zilberman.

2013. "Agricultural Biotechnology: Economics, Environment, Ethics, and the Future." *Annual Review of Environment and Resources* 38: 249–79.

Bourguet, Denis, Marion Desquilbet, Stéphane Lemarié. 2005. "Regulating Insect Resistance Management: The Case of Non-*Bt* Corn Refuges in the US. *Journal of Environmental Management* 76(3): 210–20.

Brookes, Graham, and Peter Barfoot. 2006. "GM crops: The First Ten Years–Global Socio-economic and Environmental Impacts." ISAAA Brief no. 36, International Service for the Acquisition of Agri-biotech Applications. Ithica, NY: ISAA.

Brookes, Graham, and Peter Barfoot. 2010. "Global Impact of Biotech Crops: Environmental Effects, 1996–2008." *Journal of Agrobiotechnology Management and Economics* 13(1): Article 6.

Brookes, Graham, and Peter Barfoot. 2012. "Global Impact of Biotech Crops: Environmental Effects, 1996–2010." *GM Crops Food* 3(2) 129–37.

Cerdeira, Antonio L., and Stephen O. Duke. 2006. "The Current Status and Environmental Impacts of Glyphosate-Resistant Crops." *Journal of Environmental Quality* 35(5): 1633–58.

Codex Alimentarius Commission. 2003a. "Draft Principles for the Risk Analysis of Foods Derived from Modern Biotechnology, at Step 8 of the Elaboration Procedure." ALINORM 03/34, Appendix II. Codex Alimentarius Commission, Joint FAO/WHO Food Standards Programme, Food and Agriculture Organisation, Rome.

Codex Alimentarius Commission. 2003b. "Draft Guideline for the Conduct of Food Safety Assessment of Foods Derived from Recombinant-DNA Plants, at Step 8 of the Elaboration Procedure." ALINORM 03/34, Appendix III. Codex Alimentarius Commission, Joint FAO/WHO Food Standards Programme, Food and Agriculture Organisation, Rome.

Colson, Gregory J., Wallace E. Huffman, and Matthew C. Rousu. 2011. "Improving the Nutrient Content of Food through Genetic Modification: Evidence from Experimental Auctions on Consumer Acceptance." *Journal of Agricultural and Resource Economics* 36(2): 343–64.

Current Research Information System (CRIS). 2012. Selected CRIS Funding Summaries, USDA, SAES and Other Institutions by Knowledge Area. http://cris.nifa.usda.gov/crisfin/2011/11tabled.pdf.

Crost Benjamin, Bhavani Shankar, Richard Bennett, Stephen Morse. 2007. "Bias from Farmer Self-Selection in Genetically Modified Crop Productivity Estimates: Evidence from Indian Data." *Journal of Agricultural Economics* 58(1): 24–36.

de Gorter, Harry, and David Zilberman. 1990. "On the Political Economy of Public Good Inputs in Agriculture." *American Journal of Agricultural Economics* 72(1): 131–37.

Duke, Stephen O., and Stephen B. Powles. 2009. "Glyphosate-Resistant Crops and Weeds: Now and in the Future." *AgBioForum* 12(3–4): Article 10.

Duvick, Donald N., J. S. C. Smith, and M. Cooper. 2004. "Changes in Performance, Parentage, and Genetic Diversity of Successful Corn Hybrids, 1930–2000." In *Corn: Origin, History, Technology, and Production*, edited by C. Wayne Smith, Javier Betrán, and E. C. A. Runge, 65–97. Hoboken: John Wiley & Sons.

European Food Safety Authority (EFSA). 2004. "Opinion of the Scientific Panel on Genetically Modified Organisms on the Use of Antibiotic Resistance Genes as Marker Genes in Genetically Modified Plants." *EFSA Journal* 48: 1–18.

European Commission Eurobarometer 52.1. 2000. "The Europeans and Biotechnology." Report by INRA (Europe)—ECOSA on behalf of Directorate-General for Research, Directorate B—Quality of Life and Management of Living Resources Programme. http://ec.europa.eu/research/pdf/eurobarometer-en.pdf.

European Commission Eurobarometer 73.1. 2010. "Biotechnology." Survey coordinated by Directorate General Research, conducted by TNS Opinion & Social on request of the European Commission. http://ec.europa.eu/public_opinion/archives/ebs/ebs_341_en.pdf.

Fernandez-Cornejo, Jorge, and William McBride. 2002. "Adoption of Bioengineered Crops." Agricultural Economic Report AER-810, United States Department of Agriculture Economic Research Service.

Foster, Kenneth R., Paolo Vecchia, and Michael H. Repacholi. 2000. "Science and the Precautionary Principle." *Science* 288(5468): 979–81.

Godfray, H. Charles J., John R. Beddington, Ian R. Crute, Lawrence Haddad, David Lawrence, James F. Muir, Jules Pretty, Sherman Robinson, Sandy M. Thomas, and Camilla Toulmin. 2010. "Food Security: The Challenge of Feeding 9 Billion People." *Science* 327(5967): 812–18.

Goodman, Richard E., Stefan Vieths, Hugh A. Sampson, David Hill, Motohiro Ebisawa, Steve L. Taylor, and Ronald van Ree. 2008. "Allergenicity Assessment of Genetically Modified Crops: What Makes Sense?" *Nature Biotechnology* 26(1): 73–81.

Graff, Gregory D., Gal Hochman, and David Zilberman. 2009. "The Political Economy of Agricultural Biotechnology Policies." *AgBioForum* 12(1): Article 4.

Graff, Gregory D., and David Zilberman. 2001. "An Intellectual Property Clearinghouse for Agricultural Biotechnology." *Nature Biotechnology* 19(12): 1179–80.

Graff, Gregirt D., David Zilberman, and Alan B. Bennett. 2009. "The Contraction of Agbiotech Product Quality Innovation." *Nature Biotechnology* 27(8): 702–704.

Graff, Gregory D., Susan E. Cullen, Kent J. Bradford, David Zilberman, and Alan B. Bennett. 2003. "The Public–Private Structure of Intellectual Property Ownership in Agricultural Biotechnology." *Nature Biotechnology* 21(9): 989–95.

Gouse, Marnus, Carl Pray, David Schimmelpfennig. 2004. "The Distribution of Benefits from Bt Cotton Adoption in South Africa." *AgBioForum* 7(4): 187–194.

Hare, Peter D., and Nam-Hai Chua. 2002. "Excision of Selectable Marker Genes from Transgenic Plants." *Nature Biotechnology* 20(6): 575–80.

Hoban, Thomas J. 1998. "Trends in Consumer Attitudes about Agricultural Biotechnology." *AgBioForum* 1(1): Article 2.

Huang, Jikun, Carl Pray, and Scott Rozelle. 2002. "Enhancing the Crops to Feed the Poor." *Nature* 418(6898): 678–84.

Huang, Jikun, Ruifa Hu, Scott Rozelle, and Carl Pray. 2005. "Insect-Resistant GM Rice in Farmers' Fields: Assessing Productivity and Health Effects in China." *Science* 308(5722): 688–90.

James, Clive. 2011. "Global Status of Commercialized Biotech/GM crops: 2011." ISAAA Brief 43-2011, International Service for the Acquisition of Agri-biotech Applications, Ithaca, New York. http://www.isaaa.org/resources/publications/briefs/43/.

Jasa, Paul J. 2000. "Conservation Tillage Systems." Paper presented at the International Symposium on Conservation Tillage, in Mazatlan, Mexico, January 24–27. http://agecon.okstate.edu/isct/labranza/jasa/tillagesys.doc.

Just, Richard E., Julian M. Alston, and David Zilberman, eds. 2006. *Regulating Agricultural Biotechnology: Economics and Policy.* New York: Springer.

Kathage, Jonas, and Matin Qaim. 2012. "Economic Impacts and Impact Dynamics of Bt *(Bacillus thuringiensis)* Cotton in India." *PNAS* 109(29): 11652–56.

Kiesel, Kristin, Jill J. McCluskey, and Sofia B. Villas-Boas. 2011. "Nutritional Labeling and Consumer Choices." *Annual Review of Resource Economics* 3(1): 141–58.

Koo, Bonwoo, Carol Nottenburg, and Philip G. Pardey. 2004. "Plants and Intellectual Property: An International Appraisal." *Science* 306(5700): 1295–97.

Lapan, Harvey E., and GianCarlo Moschini. 2004. "Innovation and Trade with Endogenous Market Failure: The Case of Genetically Modified Products." *American Journal of Agricultural Economics* 86(3): 634–48.

Lichtenberg, Erik, and David Zilberman. 1986. "The Econometrics of Damage Control: Why Specification Matters." *American Journal of Agricultural Economics* 68(2): 262–73.

Lusk, Jayson L., Mustafa Jamal, Lauren Kurlander, Maud Roucan, and Lesley Taulman. 2005. "A Meta-analysis of Genetically Modified Food Valuation Studies." *Journal of Agricultural and Resource Economics* 30(1): 28–44.

Lynas, Mark. 2013. "Lecture to Oxford Farming Conference, 3 January 2013." Available from http://www.marklynas.org/2013/01/lecture-to-oxford-farming-conference-3-january-2013/.

Malik, J., G. Barry, and G. Kishore. 1989. "The Herbicide Glyphosate." *BioFactors* 2(1): 17.

Marchant, Robert. 2001. "From the Test Tube to the Table." *EMBO Reports* 2(5): 354–57.

Marris, Claire. 2001. "Public Views on GMOs: Deconstructing the Myths." *EMBO Reports* 2(7): 545–48.

McHughen, Alan. 2007. "Public Perceptions of Biotechnology." *Biotechnology Journal* 2(9): 1105–11.

Mitchell, Jeffrey P., Daniel S. Munk, Bob Prys, Karen K. Klonsky, Jon F. Wroble, and Richard L. De Moura. 2006. "Conservation Tillage Production Systems Compared in San Joaquin Valley Cotton." *California Agriculture* 60(3): 140–45.

Mueller, Thomas C., Paul D. Mitchell, Bryan G. Young, and A. Stanley Culpepper. 2005. "Proactive versus Reactive Management of Glyphosate-Resistant or Tolerant Weeds 1." *Weed Technology* 19(4): 924–33.

National Research Council (NRC) Committee on Genetically Modified Pest-Protected Plants. 2000. *Genetically Modified Pest-Protected Plants: Science and Regulation.* Washington, DC: National Academies Press. http:// www.nap.edu/books/0309069300/html/.

National Research Council (NRC) Committee on Risk Assessment of Hazardous Air Pollutants. 1994. *Science and Judgment in Risk Assessment.* Washington, DC: National Academy Press.

National Research Council (NRC) Committee on Risk Characterization, Commission on Behavioral and Social Sciences and Education. 1996. *Understanding Risk: Informing Decisions in a Democratic Society,* edited by Paul C. Stern and Harvey V. Fineberg. Washington, DC: National Academy Press.

National Research Council (NRC) Committee on the Impact of Biotechnology on Farm-Level Economics and Sustainability. 2010. *Impact of Genetically Engineered Crops on Farm Sustainability in the United States.* Washington, DC: National Academies Press.

National Research Council (NRC) Committee

on Scientific Evaluation of the Introduction of Genetically Modified Microorganisms and Plants into the Environment. 1989. *Field Testing Genetically Modified Organisms: Framework for Decisions.* Washington, DC: National Academies Press.

Paarlberg, Robert. 2008. *Starved for Science: How Biotechnology is Being Kept out of Africa.* Cambridge: Harvard University Press.

Paarlberg, Robert. 2010. "GMO Foods and Crops: Africa's Choice." *New Biotechnology* 27(5): 609–13.

Pew Charitable Trusts. 2001. "Pew Initiative on Food and Biotechnology." http://www.pewhealth.org/projects/pew-initiative-on-food-and-biotechnology-85899367237.

Piggott, Nicholas E., and Michele C. Marra. 2008. "Biotechnology Adoption over Time in the Presence of Non-Pecuniary Characteristics that Directly Affect Utility: A Derived Demand Approach." *AgBioForum* 11(1): Article 6.

Pollack, Andrew. 2013. "In a Bean, a Boon to Biotech." *New York Times*, November 15. http://www.nytimes.com/2013/11/16/business/in-a-bean-a-boon-to-biotech.html?_r=0.

Pollan, Michael. 2001. *The Botany of Desire: A Plant's Eye View of the World.* New York: Random House Trade Paperbacks.

Potrykus, Ingo. 2010. "Regulation Must Be Revolutionized." *Nature* 466(7306): 561.

Qaim, Matin. 2009. "The Economics of Genetically Modified Crops." *Annual Review of Resource Economics* 1: 665–94.

Qaim, Matin, and David Zilberman. 2003. "Yield Effects of Genetically Modified Crops in Developing Countries." *Science* 299(5608): 900–902.

Ronald, Pamela. 2011. "Plant Genetics, Sustainable Agriculture and Global Food Security." *Genetics* 188(1): 11–20.

Sanders, Larry D. 2000. "The Economics of Conservation and Conservation Tillage." Paper presented at the International Symposium on Conservation Tillage, Mazatlan, Mexico, January 24–27. http://agecon.okstate.edu/isct/labranza/sanders/mazecon00.doc.

Sexton, Steven, and David Zilberman. 2011. "Land for Food and Fuel Production: The Role of Agricultural Biotechnology." Chap. 8 in *The Intended and Unintended Effects of U.S. Agricultural and Biotechnology Policies*, edited by Joshua S. Graff Zivin, and Jeffrey M. Perloff. NBER, University of Chicago Press.

Shaner, Dale L. 2000. "The Impact of Glyphosate-tolerant Crops on the Use of Other Herbicides and on Resistance Management." *Pest Management Science* 56(4): 320–26.

St. Clair, Jeffrey, and Joshua Frank. Forthcoming. *Green Scare: the New War on Environmentalism.* Book chapters are available at the bottom of this webpage: http://thewrongkindofgreen.wordpress.com/2011/05/03/from-greenpeace-to-greenwash/.

Stewart, C. Neal, Matthew D. Halfhill, and Suzanne I. Warwick. 2003. "Transgene Introgression from Genetically Modified Crops to Their Wild Relatives." *Nature Reviews Genetics* 4(10): 806–17.

Tabashnik, Bruce E., J. B. J. Van Rensburg, and Yves Carrière. 2009. "Field-evolved Insect Resistance to *Bt* Crops: Definition, Theory, and Data." *Journal of Economic Entomology* 102(6): 2011–25.

Taylor, Steve L. 2006. "Review of the Development of Methodology for Evaluating the Human Allergenic Potential of Novel Proteins." *Molecular Nutrition & Food Research* 50(7): 604–609.

Trigo, Eduardo J., and Cap, Eugenio J. 2003. "The Impact of the Introduction of Transgenic Crops in Argentinean Agriculture. *AgBioForum* 6(3): 87–94.

US Department of Agriculture–National Resources Conservation Service (USDA-NRCS). 2008. "Energy Consumption Awareness Tool: Tillage." Washington, DC. http://ecat.sc.egov.usda.gov/.

Warwick, Suzanne I., Hugh J. Beckie, and Linda M. Hall. 2009. "Gene Flow, Invasiveness, and Ecological Impact of Genetically Modified Crops." *Annals of the New York Academy of Sciences* 1168(1): 72–99.

Wolfenbarger, Laressa L., and Paul R. Phifer. 2000. "The Ecological Risks and Benefits of Genetically Engineered Plants." *Science* 290(5499): 2088–93.

Wu, Felicia. 2006. "Mycotoxin Reduction in Bt Corn: Potential Economic, Health, and Regulatory Impacts." *Transgenic Research* 15(3): 277–89.

Yorobe, Jose M., Jr., and Quicoy, C. B. 2006. "Economic Impact of Bt Corn in the Philippines." *Journal of Philippine Agricultural Scientist* 89(3): 258–67.

Zilberman, David, Holly Ameden, and Matin Qaim. 2007. "The Impact of Agricultural Biotechnology on Yields, Risks, and Biodiversity in Low-income Countries." *Journal of Development Studies* 43(1): 63–78.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 121–146

Agriculture in the Global Economy†

Julian M. Alston and Philip G. Pardey

Agriculture is diverse and full of contradictions. The sector accounts for a comparatively small share of the global economy, but remains central to the lives of a great many people. In 2012, of the world's 7.1 billion people, an estimated 1.3 billion (19 percent) were directly engaged in farming, but agriculture (including the relatively small hunting/fishing and forestry sectors) represented just 2.8 percent of overall income (World Bank 2012). However, in today's middle- and low-income countries, where most of the world's farmers are to be found, agriculture accounts for a much greater share of national income and employment—for instance, in India, agriculture represents 18 percent of national income and 54 percent of employment.[1]

Looking beyond direct employment, in 2010 about 2.6 billion people around the world depended on agriculture for their livelihoods, either as actively engaged workers or as dependents, while about half of the world's population lived in rural

■ *Julian Alston is Professor of Agricultural and Resource Economics and Director of the Robert Mondavi Institute Center for Wine Economics, both at the University of California, Davis, Davis, California. He is also a member of the Giannini Foundation of Agricultural Economics. Philip Pardey is Professor of Science and Technology Policy and Director of the International Science and Technology Practice and Policy (InSTePP) Center, both in the Department of Applied Economics, University of Minnesota, St. Paul, Minnesota. Their email addresses are julian@primal.ucdavis.edu and ppardey@umn.edu. Authorship is alphabetical.*

† To access the disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.1.121 doi=10.1257/jep.28.1.121

[1] Throughout this paper we group countries by geographic region or based on per capita Gross National Income in 2010 based on the World Bank's (2012) classification, which designated countries to be low-income if their average per capita income was $1,035 or less; lower-middle-income, $1,036–$4,085; upper-middle-income, $4,086–$12,615; and high-income, $12,616 or more.

areas and, of these, about three-quarters were estimated to be living in agriculturally based households (FAOSTAT 2013). Agriculture supplies much more than food for direct human consumption: it produces significant amounts of feed (for livestock), fuel (for transportation, energy production, including household kitchen fires), fiber (for clothing), and, increasingly, agricultural biomass used to produce a host of industrial chemical and material products.

Agricultural production takes up a lot of space—indeed, about 40 percent of the world's land area is occupied by agriculture—and the nature of the space varies in ways that are relevant for the choice of inputs, outputs, and technology, which is often very site-specific. Partly because of differences in climate and natural resource endowments around the world, farms and farming are enormously diverse—in terms of size of farms, products produced, technology used, inputs employed, farm incomes, and other economic outcomes.

Many of the world's poor still live in rural areas, although even that is changing, albeit slowly, as poverty becomes an increasingly urban phenomenon. According to Ravallion, Chen, and Sangraula (2007), in 2002 (the latest year of data they report), around three-quarters of the estimated 1.2 billion people with incomes of less than $1.08 per day (in 1993 purchasing power parity US dollars) resided in rural areas. Many are subsistence farmers, operating very small farms using very little in the way of marketable inputs other than the land they farm and their own family labor. At the same time, some farmers are relatively wealthy and earn relatively high incomes, especially in today's high-income countries where agriculture typically represents less than 2 percent of national income and employment. This diversity reflects country-specific differences in the broader economic context in which agriculture operates, as well as differences specific to agriculture, which is changing rapidly in some places.

Much of agricultural economics is concerned with understanding these patterns and how they change over time. In this article, we first look at how the high-income countries like the United States represent a declining share of global agricultural output while middle-income countries like China, India, Brazil, and Indonesia represent a rising share. We look at the differing patterns of agricultural inputs across countries and the divergent productivity paths taken by their agricultural sectors. We then look at productivity more closely and at the evidence that the global rate of agricultural productivity growth is declining—with potentially serious prospects for the price and availability of food for the poorest people in the world.

In thinking about future productivity growth in agriculture, we look at patterns of agricultural research and development efforts. While high-income countries spend a lot on agricultural research relative to the value of agricultural output, much of that spending is focused on issues other than raising agricultural productivity. Meanwhile, the middle-income countries have a rising share of global spending on agricultural research, which over time should reinforce their growing dominance in global agricultural production. Indeed, the future of global agriculture—including the question of how agriculture will feed the world's growing populations with their rising per capita incomes, and what that means for the world's poorest people—is likely to be increasingly determined by developments in the agricultural sectors of middle-income economies like China, India, Brazil, and Indonesia.

The Shifting Locus of Global Agricultural Production

Over the past half-century or so, growth in agricultural output far outstripped the rise in population, with a relatively modest increase in land. In 1961, the world produced $746 billion worth of agricultural output (in 2004–2006 prices converted at purchasing power parity) on 4.5 billion hectares of land, using 768 million workers in agriculture, to feed 3 billion people. Over the subsequent half century the world's population grew by 1.69 percent per year, more than doubling from 1960 to reach 7.0 billion people in 2011. But agricultural output increased more than threefold in real terms—growing by 2.25 percent per year—to total $2.4 trillion in 2011. Farm prices in inflation-adjusted terms declined from 1960 to 2010, so this rise in the value of agricultural output represents an even larger increase in the quantity of output. Over this time, land in agriculture increased by a measurable but much more modest 0.22 percent per year to reach 4.9 billion hectares: 31.6 percent of which was in crops, the rest used for raising livestock.

Agricultural Outputs

Farming is enormously diverse around the world: in farming systems, technologies, farm sizes, the mixture of outputs produced, the types of inputs used to produce them, and input proportions. Some of these differences reflect differences in soils and climate or infrastructure that influence agricultural possibilities, while others reflect differences in the relative prices of inputs and outputs and other factors that determine comparative advantage, as well as government policies that dampen its relevance. Some places can grow bananas and pineapples, others can grow lettuce and strawberries, and some can at best graze cattle at less than one beast per square mile. As well as affecting what can be grown, and what it is economic to grow, location affects yield and quality of production, and susceptibility to pests and diseases.[2]

The growth in agricultural output has been very uneven, as shown by the changing shares of total global production in Table 1. Today's high-income countries produced 43.8 percent of total agricultural output in 1961, and although production by the high-income countries almost doubled by 2011, their share of the global total shrank to 24.6 percent. The region comprising Eastern Europe and the former Soviet Union produced 13.8 percent of global food output in 1961, but by 2011, it was producing only 6.5 percent of the global total. Conversely, the global

[2] The concern with "terroir" in wine production exemplifies the phenomenon. Within California, prices for wine grapes in 2010 ranged from an average of over $3,000 per ton in the Napa Valley (and in some instances more than $10,000 per ton) to an average of less than $300 per ton in some crush districts in the southern San Joaquin Valley, 250 miles south (Fuller and Alston 2012).

Table 1
Global Value of Production by Region, 1961 and 2011

Region	1961 Output (2005 PPP$ billion)	1961 Share (percent)	2011 Output (2005 PPP$ billion)	2011 Share (percent)
High income	327	43.8	591	24.6
Eastern Europe and former Soviet Union	103	13.8	155	6.5
Asia and Pacific	178	23.9	1,070	44.7
Latin America and Caribbean	69	9.2	307	12.8
Middle East and North Africa	28	3.7	125	5.2
Sub-Saharan Africa	42	5.6	149	6.2
World	**746**	**100**	**2,397**	**100**

Source: Authors' calculations based on FAOSTAT (2013).
Notes: Countries are grouped according to World Bank (2012) schema, which means that high-income countries are excluded from each geographical region. For example, that Asia and Pacific excludes Japan and Singapore, and Middle East and North Africa excludes Qatar and United Arab Emirates. PPP$ are purchasing power parity dollars.

share of agricultural production increased for all other regions. In particular, the Asia and Pacific region increased its share from 23.9 percent of global agricultural output in 1961 to 44.7 percent in 2011.

To a great extent, food is produced close to where it is consumed, so patterns of production broadly reflect population patterns, albeit with some exceptions for specific farm products that are shipped from other areas. For example, wheat, soybeans, and bananas are examples of commodities for which international trade is comparatively important, and rice is an example of a commodity for which international trade is comparatively thin. Like other staple food crops, much of the world's rice is produced and consumed within the same household. In addition, many farm outputs are physically heavy and perishable, and produced in places that are economically distant from markets.

Global agricultural production has been dominated for a long time by a short list of relatively large and populous countries, but the relative importance of these countries in aggregate and in production of particular farm commodities has been shifting—in particular reflecting a decline in the relative importance of the high-income countries. In 2009–2011, just ten countries accounted for 55.8 percent of the world's cropland, and five (India, the United States, the Russian Federation, China, and Brazil) had 42.1 percent of the total. In contrast, the 100 countries with the smallest shares made up only 0.78 percent of the world's cropland area. Production is even more spatially concentrated, with more than half the world's agricultural output coming from only five countries, and almost three-quarters of the total output produced by just 20 countries.

The Food and Agriculture Organization of the United Nations (FAO) offers country-level statistics on production of crops like maize (corn), wheat, rice, cassava, pulses (a group of crops that includes beans and "grain legumes"), and soybeans, as well as for livestock including dairy, beef, pork, and poultry. Using these data, China is the first-ranked country for total agricultural output with 23.0 percent of all global agricultural output by value, which it produces using 8.0 percent of the world's cropland (and 11.7 percent of the world's agricultural area, including pasture and grazing land). China is also top-ranked in output for wheat, rice, pork, all crops, and all livestock; second-ranked for maize and poultry; and third-ranked for dairy and beef. China is among the top four producers of every commodity listed above except cassava.

The United States is second-ranked for total agricultural output at 10.1 percent of global output; first-ranked for maize, soybeans, beef, and poultry; and second-ranked for the other livestock products, dairy, and pork. India is third-ranked overall at 9.9 percent of global output: first-ranked for pulses and dairy, and second-ranked for wheat and rice. Brazil is fourth-ranked overall at 6.0 percent of global agricultural output by value; second-ranked for cassava, soybeans, and beef; and third-ranked for pulses and poultry. Indonesia is fifth-ranked at 2.5 percent of global agricultural output. Thus, four of the top five countries in global agricultural output, including the top one, are not high-income countries.

Agricultural Inputs

Patterns of input use vary systematically among countries according to their stage of development as measured by per capita income, as demonstrated in Table 2. In 1961, today's high-income countries accounted for 43.8 percent of total global agricultural output but only 23.8 percent of global population, 27.3 percent of global agricultural land use, and 8.4 percent of global agricultural labor. However, the high-income countries accounted for 78.1 percent of the world's use of fertilizer and 81.1 percent of the world's stock of tractors used in agriculture. By 2010, today's high-income countries accounted for just 25.5 percent of global agricultural output, and even further reduced shares of global population, global agricultural land use, and global agricultural labor. The high-income countries have increased their use of fertilizer by 73.2 percent, but even so, their share of the global total use has shrunk considerably; they almost doubled their stock of tractors, while their share of the respective global total also shrunk. High-income agriculture continues to make significantly greater use of modern land- and labor-saving inputs compared with agriculture in middle- and especially low-income countries.

Land-labor ratios differ enormously among countries. In 2010, the quantity of arable and permanent cropland ranged from an average of 10.9 hectares per capita of agricultural population (22.2 hectares per capita of agricultural labor) in the high-income countries down to an average of 0.24 hectares per capita of agricultural population (0.45 hectares per capita of agricultural labor) in the countries of the Asia and Pacific region. These disparities have been growing over time, as land-labor ratios have been rising in the high-income countries (as well as in Eastern Europe and the former Soviet Union), but falling elsewhere as low- and middle-income countries have intensified their use of farm labor.

Table 2
Use of Agricultural Inputs, 1961 and 2010

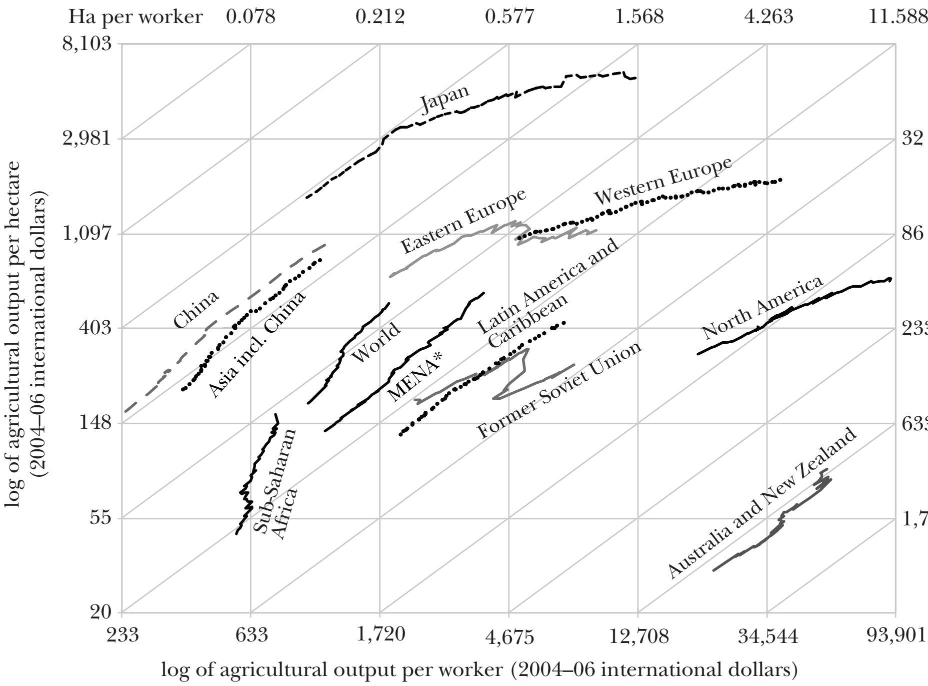
		Countries by income class				
Variable	*Unit*	*High*	*Upper middle*	*Lower middle*	*Low*	*World*
1961						
Agricultural labor	million	64.8	382.8	229.6	90.3	767.5
Agricultural land	million ha	1,107.4	1,657.3	816.0	469.9	4,050.5
Fertilizer	million ton	24.9	5.4	1.3	0.2	31.9
Tractors	million	9.2	2.0	0.1	0.1	11.3
Animal traction	million HP	16.1	61.8	63.7	11.1	152.7
Cropland per agricultural labor	ha per person	6.3	1.4	1.4	1.2	1.8
2010						
Agricultural labor	million	17.2	594.9	463.1	231.0	1,306.2
Agricultural land	million ha	1,094.1	2,009.7	973.3	562.0	4,639.2
Fertilizer	million ton	43.1	81.8	44.1	3.4	172.5
Tractors	million	16.6	7.0	3.7	0.2	27.5
Animal traction	million HP	15.0	66.1	126.1	25.1	232.3
Cropland per agricultural labor	ha per person	22.1	1.0	0.8	0.7	1.2

Source: Authors' calculations based on IFA (2013) and FAOSTAT (2013).
Notes: Countries grouped based on per capita income in 2010 according to World Bank classification (see footnote 1). Agricultural labor represents economically active population in agriculture; agricultural land is the sum of permanent pasture and harvested area; cropland is the sum of arable and permanently cropped land; fertilizer represents nitrogen, phosphate, and potash in tons of plant nutrients consumed; tractors is the number of agricultural tractors in use. According to FAOSTAT (2013) agricultural tractors "generally refers to total wheel, crawler, or track-laying type tractors and pedestrian tractors used in agriculture." Animal traction represents the stock of buffaloes, horses, asses, mules, and camels. We converted the stock of live animals to horsepower using conversion factors from Craig, Pardey, and Roseboom (1997). The abbreviation "ha" means "hectares."

Figure 1 uses the graphical technique developed by Hayami and Ruttan (1971) to track land and labor productivity movements globally and for various regions of the world. The horizontal axis is a measure of labor productivity and the vertical axis is a measure of land productivity, both plotted on a natural log scale for the period 1961–2011. Because the horizontal axis is output per worker and the vertical axis is output per hectare, it is possible to construct 45-degree lines representing constant land/labor ratios measured by hectares per worker. The lines on the graph plot a series of points from 1961 to 2011 for each region or country shown, moving generally from the lower left to the upper right over time as productivity improves.

These productivity paths exhibit clear patterns. All regions show growth in productivity over time, but at varying rates, with a longer path moving generally

Figure 1
Land and Labor Productivity by Region, 1961–2011



Source: Authors' calculations based on FAOSTAT (2013).
Notes: Diagonal lines represent constant hectare-per-agricultural worker ratios. The ratios corresponding to each diagonal line are labeled along the top and right sides of the graph in units of hectares (ha) per worker. Output is an estimate of the total value of agricultural production (spanning 192 crops and livestock commodities) expressed in 2004–06 average purchasing power parity agricultural prices from FAO (2012). Land is a measure of harvested and permanently pastured area, and labor is a head count of the total number of economically active workers in agriculture. Neither of these measures takes account of differences in land and labor quality among places and over time. Countries are grouped based on per capita income in 2010 according to World Bank classification (see footnote 1).
*Middle East and North Africa

north and east reflecting a faster rate of productivity growth over the period (see, for example, the section of the graph showing China, and Asia including China). But the starting points and the slopes of the paths vary. Over the entire period, the low- and middle-income countries as a group have much lower agricultural land-per-worker ratios than the high-income countries (that is, they are on higher diagonal hectare-per-worker lines). In 1961, both land and labor productivities in the high-income group of countries were substantially greater than the corresponding partial factor productivities in the low- and middle-income groups. Fifty years later, the disparities in measured land productivity (on the vertical axis) have become less pronounced, while the disparities in measured labor productivity (on the horizontal axis) have become more pronounced—the high-income countries produce much more output per measured unit of agricultural labor than the rest of the world.

This combination of changes suggests that productivity paths are following two distinct patterns. In the higher-income regions like Japan, Europe, the former Soviet Union, and North America, the productivity paths are comparatively flat, which means that agriculture has become less labor-intensive. In today's high-income countries, with their comparatively high wages, a major consequence of technological change and responses to relative factor prices has been to reduce the total amount of labor employed in farming and the number of people living on farms, with commensurate increases in farm sizes (Kislev and Peterson 1981). However, in other regions and for the world as a whole, the productivity patterns are relatively steep, reflecting an intensification of the use of labor relative to land in agriculture in many places. The differences in labor intensity in Figure 1 do not reflect differences in labor quality associated with differences in skill, specialization, and formal education, nor do they reflect differences in land quality. If we could adjust for these dimensions, the international divergences would be smaller: the pattern would be compressed as we revealed in Alston and Pardey (1996, pp. 156–157) for example. Still, the essential story would remain much the same.

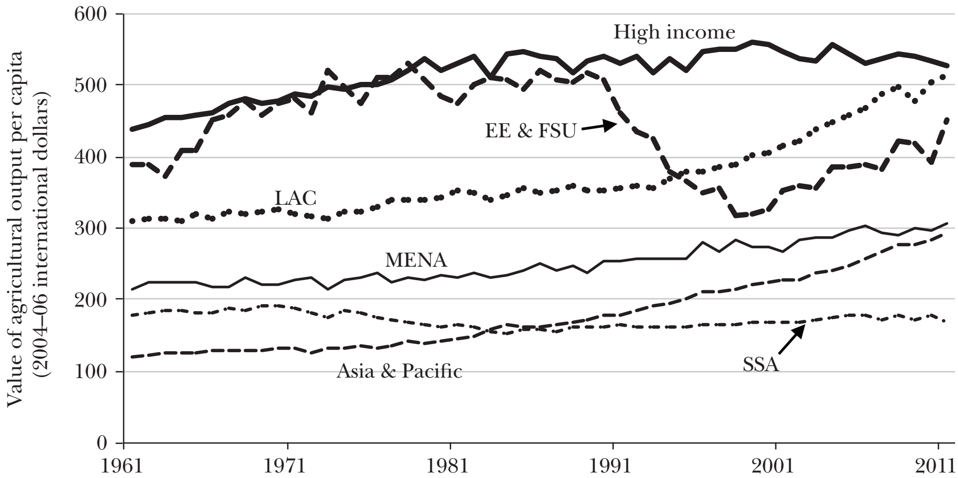
A Closer Look at Agricultural Productivity Growth

Throughout history, until relatively recently, ongoing growth in demand for food, driven both by growth in population and by generally rising per capita incomes, has been met mainly by expanding the resource base for agriculture, in particular, land. Olmstead and Rhode (2009) document the importance of biological innovation, which enabled land productivity to be sustained while land use was increasing. But during the past 100 years, and especially during the past 50 years, in most regions of the world agricultural production has been expanded mainly by increasing the output per unit of land against a relatively slowly growing land base. As noted earlier, over the period 1961 to 2011, global agricultural land use grew at a slow and shrinking rate of less than 0.22 percent per year, while population grew at about 1.7 percent per year and the real output from agriculture grew by about 2.3 percent per year.

These increases in land productivity have been accomplished by intensifying the use of "modern" inputs—in particular machinery, fertilizers, and irrigation—combined with improved genetic material and methods of production derived from organized scientific research, itself a relatively recent innovation (for a discussion in this journal, see Ruttan 2002). Along with increases in quantities of land, labor, irrigation, and fertilizer inputs, this growth reflected improvements in input quality, including new and better machines, new varieties of crops and livestock, and better-educated farmers, as well as institutional change and other changes in technology not embodied in inputs (Pardey, Alston, and Ruttan 2010; World Bank 2011).

While per capita agricultural output has grown generally, the growth has been uneven among regions and over time, as shown in Figure 2. In today's high-income

Figure 2
Per Capita Agricultural Production by Region, 1961–2011

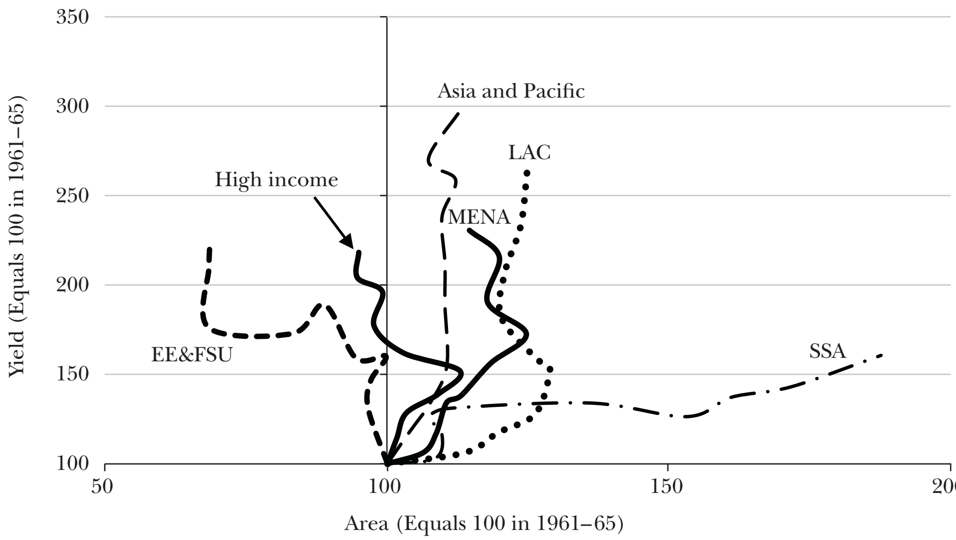


Source: Authors' calculations based on FAOSTAT (2013).
Notes: Countries are grouped according to World Bank classifications. "LAC" is Latin America and the Caribbean. "EE & FSU" is Eastern Europe the former Soviet Union. "MENA" is the Middle East and North Africa. "SSA" is sub-Saharan Africa.

group of countries, per capita agricultural output has been essentially flat since 1980. In the group of countries comprising the former Soviet Union and Eastern Europe, agricultural production collapsed following the dissolution of the Soviet Union in 1989 (Swinnen, Van Herck, and Vranken 2010) and has not yet fully recovered. However, in many of today's middle-income countries—including the regions Latin America and the Caribbean, Asia and Pacific, and Middle East and North Africa—production per capita has grown rapidly, even with reasonably rapid population growth. The picture is comparatively dismal for sub-Saharan Africa, with a decline in the real value of agricultural output per capita of 0.17 percent per year from 1961 to 2011; this decline reflects a reasonably rapid rate of growth in real output (2.6 percent per year), but it is outstripped by a more rapid growth in population (2.8 percent per year).

It is challenging to partition growth of agricultural productivity among different possible sources because the available data for many countries are very limited, especially for inputs and, within that aggregate, especially for capital inputs. Certainly the growth in output has reflected a combination of changes in the total quantity and mix of inputs as well as changes in productivity, and the importance of these different elements has varied tremendously over time as well as among countries at a point in time. For a vivid example, consider the time paths since 1961–65 of indexes of regional cereal yield as shown on the vertical axis in Figure 3 plotted against a corresponding index of land planted to cereals. During the period 1961–2010, growth in cereal production in sub-Saharan Africa was almost entirely attributable to growth in area, with very slow yield growth. In contrast, in the Asia and Pacific region and the Latin America and Caribbean region, both area and especially yield

Figure 3
Expansion in Yield versus Area of Cereal Production, 1961 to 2010



Source: Authors' calculations based on FAOSTAT (2013).
Notes: Figure 3 presents the time paths since 1961–65 of indexes of regional cereal yield as shown on the vertical axis plotted against a corresponding index of land planted to cereals. The plots represent five-year averages, starting with 1961–65, which is set to 100. The terminal value for each plot is the 2006–2010 average. Countries are grouped according to World Bank classifications. "EE&FSU" is Eastern Europe the former Soviet Union. "LAC" is Latin America and the Caribbean. "SSA" is sub-Saharan Africa. Cereals include the following commodities: barley, buckwheat, canary seed, fonio, maize, millet, mixed grain, oats, quinoa, rice, rye, sorghum, triticale, wheat, and other cereals.

grew very rapidly, while for Eastern Europe and the former Soviet Union and the high-income countries, yield increased somewhat but land area declined.

An Agricultural Productivity Slowdown

The measured growth rate of crop yields—that is, the total quantity of crop produced per unit of land area harvested per year—has been slowing generally:[3] worldwide aggregate cereal yields have been growing linearly, which implies diminishing proportional growth rates.[4] This slowdown in yields is widespread and pervasive, occurring across most geographical regions and across countries with

[3] Measurement issues abound in this context. For example, even something as apparently obvious as measuring the growth in crop yield requires care in defining the numerator (for example, harvested versus marketed grain, how to adjust for quality or types such as durum versus winter versus spring wheat), the denominator (for example, area sown versus area harvested, and how to accommodate crop failures), and the time unit of the analysis (annual versus seasonal, properly matching the numerator to the denominator, and allowing for the fact that in some cases more than one crop per year is grown on a particular piece of land).

[4] Production and area data were aggregated into geographical and income groupings prior to calculating yields. Cereals include the following: barley, buckwheat, canary seed, fonio, maize, millet, mixed grain, oats, rice, rye, sorghum, triticale, wheat, and other cereals.

high, medium, and low per capita income.[5] Our own calculations based on data from FAOSTAT (2013) find that global average maize yields grew at an average annual rate of 2.33 percent for the period 1961–1990, but then by 1.77 percent for the period 1990–2011; wheat yields grew 2.72 percent per year for 1960–1990, but 1.09 percent per year for 1990–2011; rice (paddy) yields grew 2.14 percent per year for 1961–1990, and 1.06 percent per year for 1990–2011; soybean yields grew by 1.72 percent per year for 1961–1990, but 1.21 percent per year for 1990–2011; and for cereals, yields grew by 2.35 percent per year for 1961–1990, but 1.48 percent per year for 1990–2011.

Measures of land and labor productivity growth exhibit mixed patterns among countries and over time. For the low-income countries, both of these partial productivity measures grew at rates that were consistently below the rates recorded for the middle-income group of countries, although in more recent years the average rates of growth in the low-income countries appear to have risen relative to the other groups of countries. Likewise, the growth rates of land and labor productivity appear to have risen for all other country groups relative to those for the high-income countries. Table 3 clarifies the structure of pre- and post-1990 growth rates. Taking China out of the world picture, both land and labor productivity growth rates were slower after 1990 than before. The same was true for the group of 154 countries that each had a small share of the global value of agricultural production, and which combined produced only 20 percent of the 2011 total value of output from the 183 countries excluding China. In both instances, the slowdown was more pronounced in labor productivity than land productivity.

No one disputes the evidence that growth rates of crop yields and land and labor productivity have slowed for the world as a whole excluding China. We see some evidence that also points to a slowdown in more comprehensive measures of multifactor productivity (MFP), which some call total factor productivity (TFP).[6] But the evidence on either measure of productivity is much less complete, the measures that do exist are much more open to question, and the issue is somewhat controversial.

[5] Various approaches might be used to compute average annual growth rates, as discussed by the World Bank (2013). The main alternatives are the "exponential growth rate," which is determined entirely by the starting and ending points and the "least-squares growth rate," which is obtained by regressing the natural logarithm of the measure of productivity against time. The slope coefficient from this regression is an estimate of the rate of productivity growth. Compared with the exponential growth rate, the least squares growth rate is less sensitive to starting and ending points of subperiods but more sensitive to other outliers in the sample. This method is also subject to bias from specification error, if the true path of productivity growth is not exponential, or from other failures of the linear regression model. All yield growth rates reported here were calculated using the least-squares method.

[6] Total factor productivity (TFP) is conceived as a measure of the aggregate quantum of *all* outputs divided by the aggregate quantum of *all* of the inputs used to produce those outputs. As we observe in Alston, Babcock, and Pardey (2010, p. 452): "TFP is a theoretical concept. All real-world measures omit at least some of the relevant outputs and some of the relevant inputs, and therefore it is more accurate to refer to the real-world measures as multifactor productivity (MFP) measures. Particular MFP measures differ in the extent to which they fall short of the counterpart ideal TFP measure because of methodological differences as well as differences in the consequences of incomplete coverage of the inputs and outputs."

Table 3
Land and Labor Productivity Growth before and after 1990

	Land productivity growth			Labor productivity growth		
Grouping	*1961–90*	*1990–2011*	*Difference*	*1961–90*	*1990–2011*	*Difference*
			(percent per year)			
World (184 countries)	2.04	2.22	0.18	0.93	1.71	0.79
China	2.89	4.00	1.11	2.03	4.13	2.10
World minus China						
Total (183 countries)	1.92	1.78	–0.14	1.09	0.92	–0.17
Top 29 countries	1.91	1.85	–0.06	1.15	1.26	0.12
Bottom 154 countries	1.89	1.65	–0.24	0.93	0.12	–0.81

Source: Authors' calculations based on FAOSTAT (2013).
Notes: The 29 countries that account for 80 percent of total agricultural production from the 183 countries, comprising the world excluding China, are represented in the second to the last row of the table. The last row represents the productivity of the remaining 154 out of 183 countries. All aggregates are formed as weighted averages. Difference is given as the rate for 1990–2011 minus the rate for 1961–90.

Using FAO data for 171 countries for 1961–2009, Fuglie (2012, p. 356) has undertaken an extensive analysis of agricultural productivity patterns and concludes: "[T]here does not seem to be a slowdown in sector-wide agricultural productivity growth. If anything, the growth rate in agricultural TFP [total factor productivity] accelerated in recent decades, in no small part because of rapid productivity gains achieved by developing countries, led by Brazil and China, and more recently because of a recovery of agricultural growth in the countries of the former Soviet Union and Eastern Europe." Other studies based on the same data resources have tended to find similar results (for a selection of such studies, see the chapters in Fuglie, Wang, and Ball 2012). These findings of accelerating growth in multifactor or total factor productivity are somewhat surprising given relatively constant or slowing growth in crop yields and broader partial productivity measures. We suspect that the underlying issue here lies in a host of data and measurement problems, many of which have long been discussed in the agricultural economics literature.[7] Fuglie (2012) is aware of these issues and spent considerable effort trying to address them; in particular, he notes some weaknesses of the proxy measures used to measure capital, material, or other inputs. Butzer, Mundlak, and Larson (2012) raise specific concerns about the measures of capital and their implications.

[7] Many of these data and measurement issues were identified initially by Schultz (1956) and Griliches (1963) and have been the subject of continuing efforts in the more recent literature. For instance, in Alston, Anderson, James, and Pardey (2010), we demonstrated the considerable sensitivity of their US multifactor productivity measures to choices of price weights, input quality or compositional adjustments, and measurement methods; sensitivities that are likely to be magnified in efforts to generate multifactor productivity measures on an international scale with incomplete, inconsistent, and inaccurate measures of agricultural input quantities and prices.

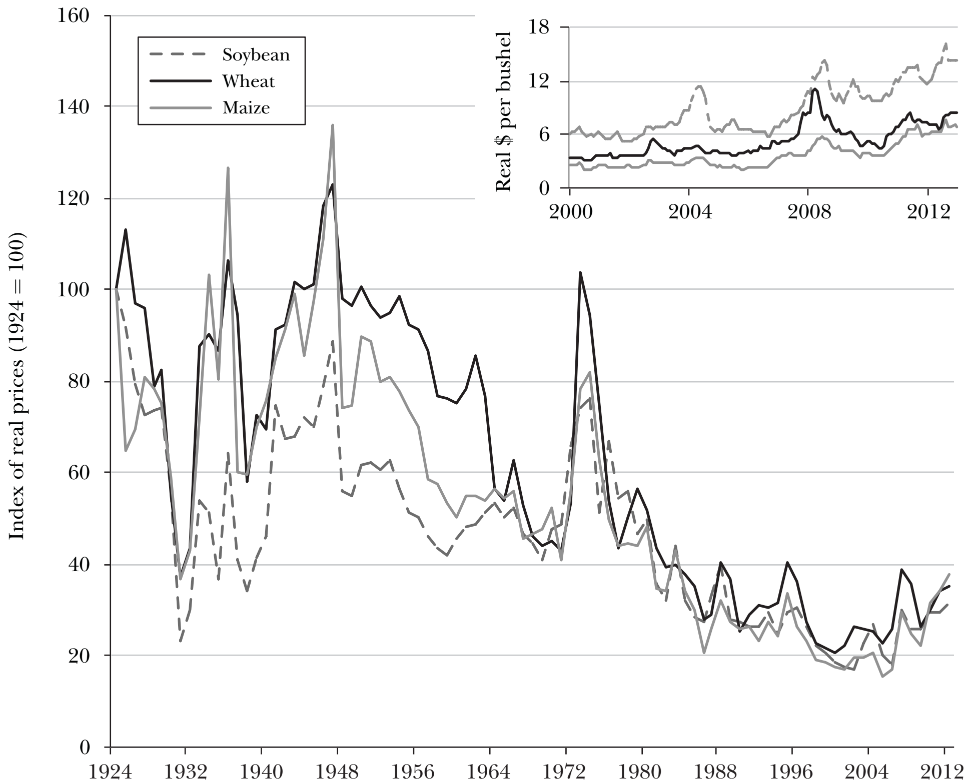
In Alston, Babcock, and Pardey (2010), we discuss the measurement issues more broadly and suggest that caution should be exercised in particular in using total factor productivity measures based on the FAO data to test for a slowdown in productivity growth because the nature of the likely index number biases is particularly damaging in that context. In some countries, much more complete and more detailed nationally sourced data permit the computation of measures of multifactor or total factor productivity that are less prone to index number biases and other measurement problems compared with studies based on FAO estimates of various agricultural inputs. For example, in Pardey, Alston, and Chan-Kang (2013), we estimated multifactor productivity growth for the US agricultural sector for 1949–2007 using the InSTePP data.[8] The results are quite different from Fuglie's.

Fuglie (2012) reported an acceleration of US agricultural productivity growth from 1.21 percent per year in the 1960s to 2.25 percent per year in 2001–2009. In contrast, in Pardey, Alston, and Chan-Kang (2012), we identified a pronounced slowdown in US agricultural productivity growth, a pattern that is consistent with the observed slowdown in US crop yield growth and other measures of US partial factor productivity growth. Likewise, using a completely different set of measures of aggregate US total factor productivity, Ball, Schimmelpfenning, and Wang (2013) also report evidence of a slowdown in agricultural productivity growth, albeit with different timing. In earlier work using US state-level InSTePP data 1949–2002, in Alston, Andersen, James, and Pardey (2010), we conducted a variety of statistical tests that comprehensively suggest that agricultural multifactor productivity growth slowed substantially after 1990, returning to a longer-term trend rate in the range of 1 percent per annum.

What should we believe about the pattern of agricultural productivity? We conclude that agricultural productivity growth generally has slowed in much of the world, both because it is the story told by crop yields and other partial productivity measures and because it is the story told by the limited set of more reliable estimates of multifactor productivity based on the detailed national-level data. Index number biases from using incomplete and ill-measured input quantities (especially for capital and materials inputs), failure to adjust for input quality improvements (especially for labor and capital), or the use of inappropriate input shares might account for the discrepancies. Thus, we have reservations about the use of measures of total factor productivity estimated using the FAO data, particularly in relation to the question of a slowdown in productivity growth over time. But the timing, extent, and likely persistence of the slowdown in agricultural productivity growth around the world are lively research topics, with crucial implications for how the food supply can expand to meet the growth in worldwide demand stemming from future increases in population and per capita income.

[8] These represent an updated version of the data we used in Alston, Andersen, James, and Pardey (2010). Details on data sources and methods used to compute the indices can be found at http://faculty.apec.umn.edu/ppardey/data-agprod.html.

Figure 4
Real US Prices of Maize, Soybeans, and Wheat, 1924–2012



Source: Beddow and Pardey (2013) and US Department of Agriculture, National Agricultural Statistics Service (2013).
Notes: The main figure plots annual average US prices received by farmers deflated by the US implicit price deflator for GDP for the period 1924 to 2012, indexed to 1924 = 100. The inset plots are real monthly prices for the period 2000–2012, expressed in 2012 values.

Prices, Productivity, Policy

The past patterns of growth in productivity and production have had important implications for food security and for poverty especially through their consequences for prices (Alston, Beddow, and Pardey 2009). During the 20th century, growth in agricultural supply more than kept pace with growth in agricultural demand—which was fueled by growth in population, income per capita, and more recently as a feedstock for biofuels—to the extent that farm commodity prices trended down significantly relative to prices generally in the economy, as illustrated for maize, soybeans, and wheat in Figure 4. But this trend has been slowing (as can be seen in Figure 4) if we set aside the period around the price spike of the early 1970s. In fact, prices have trended up since around 2000, a trend that was exacerbated by the commodity price spikes of 2007 and since, that were associated with a confluence of influences, including the rise of biofuels—encouraged by ethanol blending mandates and subsidies in the United States and elsewhere, in conjunction with high oil prices (for example, de Gorter and Just 2009; Lapan and Moschini 2012; Wright, this issue). However, the recent price rise is more properly seen as simply the last step in a progressive slowing of the rate of decline of real commodity prices over the 40 years since 1970, which is consistent with the evidence of a slowdown in growth of primal measures of agricultural productivity. Long-term trends in real prices for farm commodities are indicative of the evolving path of worldwide agricultural productivity, since the ratio of an index of input prices to an index of output prices is itself a dual measure of total factor productivity—though less useful for that purpose in short-run comparisons when relative price movements might not reflect long-run equilibrium relationships (Jorgenson and Griliches 1967).

The rise in agricultural prices in recent years reminded policymakers that agriculture plays several roles relative to food security and poverty in high- and low-income countries alike. Among the world's very poor, subsistence farmers are especially vulnerable to the effects of weather and other supply shocks that affect both their individual production and the prices they face. However, increasingly the world's poor are not farmers; nevertheless, their real incomes depend directly on the availability and price of food. Changes in agricultural technology that increase productivity or reduce susceptibility of production to weather and other shocks can contribute to the welfare of the farm and nonfarm poor by increasing abundance and lowering the real cost of food—in effect, by shifting the consumption-based poverty line—as well as by increasing net farm incomes and by reducing the variability of prices, production, and income (Alston, Martin, and Pardey forthcoming). These are all elements of the food security question that is an increasingly important part of the agricultural policy agenda. Agricultural innovation, driven increasingly by organized agricultural science, has played a large and growing role in driving agricultural development and the balance of the world food equation—as acknowledged, for example, when Norman Borlaug was awarded the Nobel Peace Prize in 1970 for his role in developing "Green Revolution" technologies.

Agricultural Innovation

The domestication of crops initially involved saving seed from one season for planting in subsequent years. Agricultural innovation began when farmers purposefully selected seeds from successful crop varieties and, by repeated selection over many years, began adapting crop genetics to the environment in which the crop was grown. Scientifically bred varieties of crops and livestock have a history of less than 200 years. Collective, more organized, and increasingly scientific forms of agricultural innovation by way of publicly funded research and development undertaken in universities and, eventually, government research agencies took hold in the early 19th century, initially in Germany, then spreading to North and South America and Australasia and, eventually, to today's developing world (for example, Ruttan 1982; Grantham 1984). Along with genetic innovations in crops, this growing body of scientific knowledge fostered innovations in pest and disease management, animal husbandry, and the like, which accompanied "labor-saving" innovations that augmented and replaced human labor with wind- and watermills, and livestock draft power, eventually to be replaced with tractors and other machines. While technical changes arising from investments in agricultural productivity are not synonymous with changes in multifactor productivity, available empirical evidence strongly supports the notion that an accumulation of past and present research spending plays a big part in stimulating present and future productivity growth in agriculture.

Role of Public and Private Agricultural Research

Some distinctive features of agriculture have shaped the division of labor between the private and public sectors in agricultural science in ways that differ from industrial science more generally. First, agriculture continues to be an atomistic sector, such that individual farm firms have had attenuated incentives to invest in many types of innovations, leaving a larger potential role for government than in industrial research generally. Second, the demands for farming technologies reflect the biological nature of production and the site-specific nature of the production environment. Hence, as well as demanding innovations that save costs or enhance product quality generally, farmers demand innovations that reduce the susceptibility of production to uncontrolled factors. In addition, they have to deal with biological obsolescence: continuing innovation is necessary simply to maintain yields as the climate changes and pests, diseases, weeds, and other aspects of the environment co-evolve.

Reflecting this context, the role of publicly funded and performed agricultural and food research rose in tandem with the growth in private research, especially among the high-income countries during the latter half of the twentieth century. Several broad institutional and market changes contributed to this evolution, reflecting the peculiar features of agriculture and its role in the economy. One contributing factor was general scientific developments, which led to particular innovations such as the development and commercialization of hybrid seeds, beginning with corn in the United States in the 1920s—but now hybrid seeds are common in other crops like canola, rice, and numerous vegetable crops such as cabbage, broccoli, and watermelon. A second critical influence was an expansion in the scope of intellectual property protection to encompass plant and animal innovations. In particular, this reform spurred the rapid expansion of private investment to develop private varieties of crops, including the genetically modified varieties of soybeans, maize, cotton, canola, and other crops that have been widely adopted in the largest agricultural countries since the mid-1990s. A third major influence was the increase in the share of food and beverages being consumed away from the home or prepared and packaged in more convenient forms, which induced innovative activity to service these changing consumer demands. For example, between 1940 and 2011, food consumed away from home rose from 19.7 percent to 48.7 percent of total US food expenditures (USDA, ERS 2012). The various changes in science, institutions (including intellectual property and regulatory realities, as well as changes in the law), and markets have tended to increase the private appropriability of the benefits arising from agricultural research, thus spurring increased attention from the private sector—especially into mechanical and chemical technologies, and more recently, new varieties for some crops, where intellectual property protection and market incentives are strong—primarily in the higher-income countries.

For the world as a whole, governments continue to outspend the private sector in agricultural research. In 2009, a total of about $35 billion (in 2005 dollars converted at purchasing power parity exchange rates) was spent on public sector agricultural and food research worldwide (Pardey, Chan-Kang, and Dehmer forthcoming). While our empirical handle on private investments in food and agricultural research is far from certain, the available evidence indicates that spending by the private sector on food and agricultural research is in the range of $20–22 billion per year. The lion's share of that research, around 90 percent, took place in the high-income countries and, for the high-income countries at least, almost one-half of that research was concerned with producing off-farm innovations, primarily those related to food processing.[9] In 2009, the United States accounted for around one-third of overall public and private spending by high-income countries on food and agricultural research.

Knowledge and Technology Spillovers

The potential for spillovers of agricultural technologies is mitigated by differences in climate and other aspects of the natural resource stocks that govern agricultural potential. Even so, spatial movement of agricultural technologies has played an important role in the development of agriculture. Before the modern scientific age, great advances in American agricultural productivity resulted from plant prospectors who imported new and improved crop varieties from foreign lands. As early as the mid-1850s, the US government employed "agricultural explorers" to scout the globe for new plant and seed material for shipping back to the United States (Ryerson 1933, p. 117). US Navy expeditions and Consular staff were also regularly used to collect new plant material from countries around the world and ship that material back to the United States (Juma 1989; Olmstead and Rhode 2007). These efforts were particularly fruitful in introducing varieties of wheat and other crops suitable for the relatively arid and harsh conditions on the Great Plains. Countries around the world made similar efforts.

Positive economic spillovers across national borders from research are still of substantial importance to global and US agriculture. For example, foreign entities accounted for 64 percent of all the plant varietal rights in the United States in 2008, compared with just 21 percent in 1984 (Pardey, Koo, Drew, Horwich, and Nottenburg 2013). By the early 1990s, about one-fifth of the total US wheat acreage was sown to varieties with ancestry developed by the International Maize and Wheat

[9] For a discussion of the range of estimates, see Pardey, Alston, and Chan-Kang (2012), where we discuss estimates from Pardey and Beintema (2001); Beintema and Stads (2008); Fuglie et al. (2011); and Beintema, Stads, Fuglie, and Heisey (2012). Pardey and Chan-Kang (forthcoming) report a public and private spending total of $37.3 million (in 2009 prices converted at purchasing power parity) for the high-income countries in 2009, 52 percent of which was performed by the private sector.

Improvement Center (CIMMYT), located in Mexico, and in 1993 virtually all the California spring wheat crop was grown with varieties from CIMMYT or with CIMMYT-based ancestors (Pardey, Alston, Christian, and Fan 1996). The US reliance on wheat varieties from CIMMYT and elsewhere in the world has persisted (Pardey and Beddow 2013, Box 3).

Agricultural research spillovers go well beyond wheat varieties and national borders. In a comprehensive study of the state-by-state returns to *all* the state-specific investments in agricultural research performed in the United States (Alston, Andersen, James, and Pardey 2009), we estimated that, on average, one-third of the economic benefits from research-induced productivity gains in agriculture in each state were attributable to spill-ins from research done in other states or by the federal government. In Alston, Anderson, James, and Pardey (2011), we report that on average across the states, an incremental dollar invested in research carried out by a state agricultural experiment station in the United States returned benefits to that state over 50 years with a real present value of $21, and a value of $32 to the nation as a whole if spillover benefits to other states were added to the own-state benefits.

Formal international collaboration in the development and diffusion of agricultural innovations, designed to enhance the international spillovers of technologies and techniques, crystalized in the first half of the twentieth century. There has been a recent revival in international agricultural research, including research undertaken by the international research centers collectively known as the Consultative Group on International Agriculture (CGIAR). The CGIAR was founded in 1971 as a collective funding instrument to support agricultural research with an explicit international intent. The system spent $20 million (nominal prices) in 1971, growing to $643 million in 2010 (for details, see Pardey and Beddow 2013; Ozgediz 2012; Wright 2012). However, the donor pressures on that system appear to be placing increased emphasis on shorter-term, more development-oriented efforts at the expense of research with longer-run payoffs and larger spillover potentials.

The potential for spillover benefits from an increase in private participation in agricultural research is also uncertain. On the one hand, the business model of multinational companies would support the increased international movement of innovations in the food and agricultural sectors. On the other hand, the cross-country flows of agricultural technologies will be limited by regulatory restrictions, lack of effective legal processes and intellectual property protection, and, especially in many of the world's poorer countries whose economies are still heavily reliant on agriculture, a preponderance of comparatively small, fragmented, and costly-to-service farms.

An Agricultural Research Spending Slowdown

In the high-income countries, in spite of compelling evidence of high rates of return and a significant productivity slowdown, public support for agricultural science has broadly waned.[10] In 1960, $5.4 billion was spent on public food and

[10] Reviewing the rates-of-return to agricultural research literature over the past 50 years, Rao, Hurley, and Pardey (2012) compiled 2,186 evaluations from 359 separate published studies and report an average

agricultural research, and today's high-income countries accounted for 56 percent of the world's total. Almost 50 years later, in 2009, that high-income country share had dropped to 48 percent, with the US share dropping from 21 percent to just 13 percent of global public spending on food and agricultural research over the same period (Pardey, Alston, and Chan-Kang 2013). Real rates of annual public research spending have begun to decline in many countries, including the United States (Pardey, Alston, and Chan-Kang 2013).

Moreover, of the amounts being spent on "agricultural science" in high-income countries, an ever-increasing share is being directed towards off-farm issues—such as health and nutrition, food safety, biofuels technology, and the environment—leaving less for research directed at maintaining or increasing farm productivity. In the United States, around 65 percent of the public research conducted by the land grant universities in 1976 was classified as farm productivity research, dropping to just 56 percent by 2009 (Pardey, Alston, and Chan-Kang 2013). Public sector research capacity in the agricultural sciences in many (especially high-income) countries has been run down over decades, infrastructure has depreciated, and the majority of the scientists focused on this research in many countries are close to retirement age.

On the salutary side, agricultural research is on the rise in large, populous middle-income countries: Brazil, India, and China together now provide 31.1 percent of the world's public agricultural research. These countries have among the largest total numbers of farmers and the "food-poor," whose lives can be very substantially improved through agricultural innovation leading to more abundant and cheaper food. By 2007–09, China was spending more than any other country on public sector agricultural research and development, with a budget of $5.8 billion per year in 2007–2009, greater than that of the United States at $4.5 billion per year. But we also see continuation of a growing global divide, with the world's poorest countries falling even farther behind. Notably, the nations of sub-Saharan Africa spent just 6 percent of the world's total public sector agricultural research in 2009, down from a 10 percent share in 1960.

Over the past 50 years, high-income countries progressed steadily towards an ever more research-intensive mode of agricultural production. Compared with just 56 cents for every $100 of agricultural output in 1960, these countries invested an average of $3.59 into public agricultural research per $100 of agricultural output in 2009. Agricultural research intensity has increased in spite of a slowdown in the rate of growth of this spending—reflecting an even more pronounced slowdown in the rate of growth of agricultural output in these countries. As productivity increases, an increasing share of research is required to sustain past productivity gains given the co-evolutionary pressures of pests and diseases that act to undercut the efficacy of

internal rate of return of 49.4 percent per year (after eliminating extreme outliers and a median of 40.7 percent per year). In Alston, Andersen, James, and Pardey (2011), we raised concerns over using internal rates of return as a summary measure of the benefits associated with a given research cost, arguing instead for a benefit–cost ratio or a "modified internal rate of return." Rao, Hurley, and Pardey (2012) report, for the studies they evaluated, an average modified internal rate of return of 11–14 percent, depending on the assumptions used when recalibrating the reported internal rates of return.

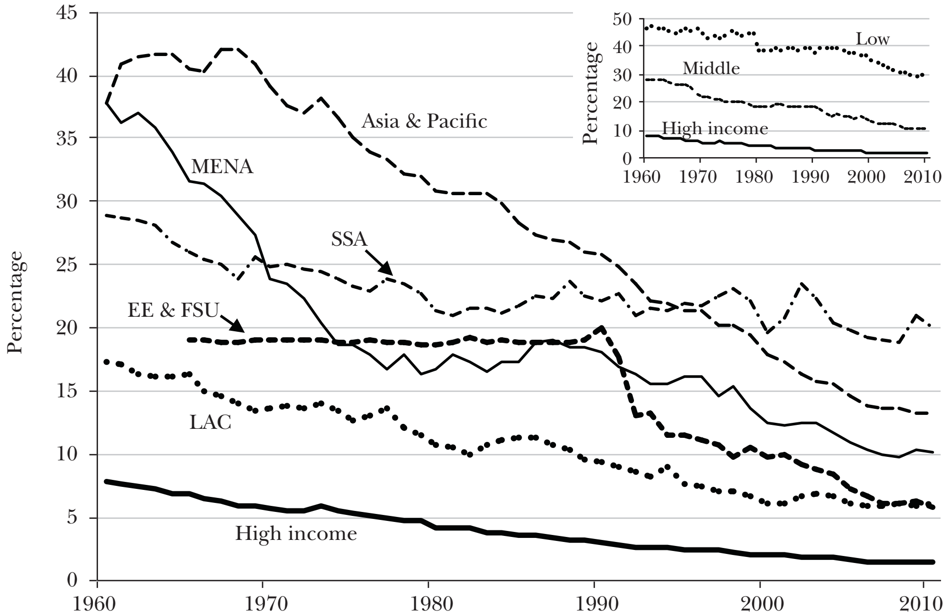
chemical, management, and biological pest-control technologies (Ruttan 1982). In contrast, the intensity with which the Asia and Pacific region invests in agricultural research has grown much more modestly from 40 cents for every $100 of agricultural output in 1960 to 54 cents in 2009. While this region has sustained growth in agricultural research spending at a comparatively rapid pace, averaging a growth rate of 5.1 percent per year since 1960, agricultural output has grown at a reasonably rapid rate of 3.8 percent per year. Thus, while the growth in spending on agricultural research and development outpaced the corresponding growth in the value of output in this region, the growth rate differentials were comparatively modest such that the region's research intensity only inched up over time, albeit increasingly so after the mid-1990s. In sub-Saharan Africa, research intensities have been slipping, especially in the past couple of decades. The same pattern is evident in the low-income countries more generally.

The Transition of Agriculture as Economies Grow

As per capita income rises in a country, the agricultural share of GDP typically falls (Timmer 2009). For example, in 2010, agriculture (again including forestry and hunting/fishing) contributed 29.3 percent of total GDP on average in countries with per capita incomes less than $1,005 (the World Bank 2010 threshold designating low-income countries); while for middle-income and high-income countries, the shares were 10.5 and 1.5 percent respectively. As Figure 5 shows, the agricultural shares of GDP have declined in every region of the world, but unevenly. In the high-income countries of the world, agriculture contributed a relatively small share of total GDP a half-century ago, and this share has dropped even further since then. Over the same time period, in the Asia and Pacific region, the agricultural share began much higher, but declined relatively rapidly, with the region's very rapid rates of economic growth. On the other hand, in sub-Saharan Africa, agriculture's share of GDP was much lower than in the Asia and Pacific region in 1960, but with its relatively slow rate of agricultural and general economic growth, agriculture's share of GDP in sub-Saharan Africa has fallen relatively slowly such that the share of GDP from agriculture is now larger than in any other region.

The cross-sectional evidence using 2005–2010 average data tells a similar story to the time-series evidence. Figure 6 plots country-specific measures of the share of labor in agriculture against GDP per capita (Figure 6A), and the share of GDP from agriculture against GDP per capita (Figure 6B). The individual bubbles represent country-specific average observations for 2005–2010. For a selection of countries, we have also plotted the time path of the share of labor in agriculture versus GDP per capita (in Figure 6A) and the share of GDP in agriculture versus GDP per capita (Figure 6B). The relationship is clearly negative in general—a larger share of total income from agriculture and of total labor engaged in agriculture is associated with a lower per capita income—though not always smooth and monotonic in the time-series plots for particular countries.

Figure 5
Agricultural GDP as Share of GDP (by Region), 1960-2010



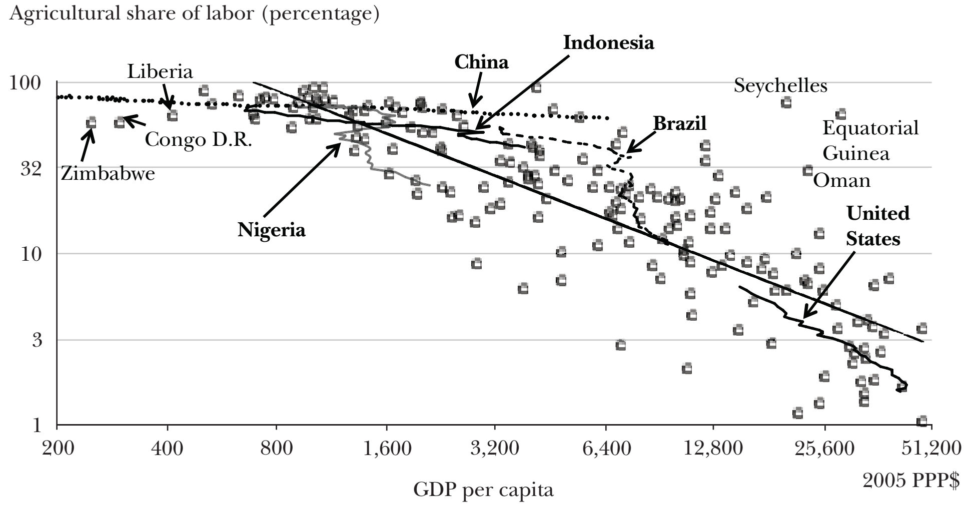
Source: Authors' calculations based on World Bank (2012) and Pardey, Chan-Kang, and Dehmer (forthcoming).
Notes: Countries are grouped according to World Bank classifications. High-income countries are excluded from each geographical region. For example, Asia & Pacific excludes Japan and Singapore; MENA (Middle East & North Africa) excludes Qatar and United Arab Emirates. High-income countries are those with 2010 GNI per capita of $12,276 or more; middle-income countries had 2010 GNI per capita between $1,006 and $12,275; and low-income countries had 2010 GNI per capita less than or equal to $1,005 (World Bank 2011, p. 389).

Around the world today can be found countries at every stage of the transition that is now largely complete in the high-income countries. In the United States, for example, the total farm population peaked at 32.5 million people, or 31.9 percent of the total US population, in 1916. Since then the US population has continued to grow while the farm population declined to 2.9 million in 2006, just 1 percent of the total population of roughly 310 million. Among today's upper-middle-income countries, China has been transforming its agricultural sector and releasing considerable labor for other employment, but nonetheless the rural and urban labor markets have not fully adjusted. Among today's lower-middle-income countries, in India 48.4 percent of the population is still agricultural and 68.7 percent earn less than $2 per day. Many low-income countries are like Mali, where over 70 percent of the population live on farms and about 80 percent earn less than $2 per day, and have not really begun the process.

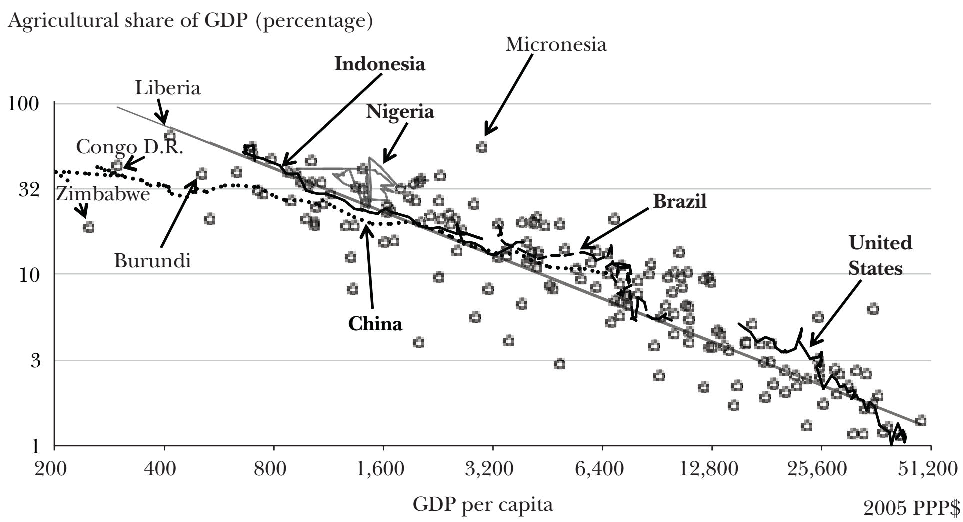
Timmer (2009) observed that the decline in the agricultural share of labor generally lagged the decline in the agricultural share of GDP, reflecting some "stickiness" of adjustments in farm labor. Fifty to 100 years ago, the same phenomenon gave rise

Figure 6
Share of Labor and GDP in Agriculture versus GDP per Capita
(logarithmic scale)

A: Share of Agricultural Labor in Total Labor



B: Share of Agricultural GDP in GDP



Source: Authors' calculations based on World Bank (2012) and FAOSTAT (2013).
Notes: The downward sloping straight line represents an OLS linear best fit. Other line plots represent the time path of annual values of the respective labor and output shares plotted against GDP per capita (both on a logarithmic scale) for the period 1961–2011 for Brazil, China, Indonesia, Nigeria, and the United States.

to the "farm problem" in the United States and other countries, a persistent problem of excess farm labor and low farm incomes that characterized agriculture in transition (for example, Gardner 1983, 2002). The coming decades may see monumental changes in structure of the farm sector in the countries that have barely begun the process of transition. But the evidence from the past suggests that this transition may not be smooth and may involve very substantial costs of adjustment (Timmer 2009).

Although it seems clear that economic growth entails a reduction in the relative importance of agriculture in an economy, the causal mechanisms in the process by which this transformation takes place are not fully clear. This transition away from agriculture seems to involve both relatively rapid growth of the nonfarm sector, and a substantial reduction in the labor intensity of farming. Much has been written about the role of innovation and economic growth in agriculture, in the first instance, as a means of priming the pump for broader economic growth by generating an economic surplus, releasing resources to and demanding inputs from the rest of the economy (for example, World Bank 2007). The hard empirical challenge is to sort out the relative roles of push from agriculture versus pull from growth in the rest of the economy, as contributors to a drift from farm work to part-time farming or nonfarm employment, in a world in which both push and pull processes are at work, synergistically, occurring in conjunction with changes in educational status, per capita incomes, and other changes, as elaborated in Timmer (2009). Sorting out these issues is a classic and still largely unresolved question at the intersection of agricultural economics and growth theory. But there can be no doubt that agricultural innovation plays a crucial role, whether in priming the pump for economic progress or as an induced innovation response to facilitate adjustment when the rest of the economy is bidding resources away from agriculture.

A Changing World Order

The past 50–100 years have witnessed dramatic changes in agricultural production and productivity, driven to a great extent by public and private investments in agricultural research, with profound implications especially for the world's poor. A trend of slowing growth, or real reductions in spending by the high-income countries on agricultural productivity enhancing research, has contributed to a slowdown in their agricultural productivity growth. But over time and among countries, the developments in agricultural production and productivity have been uneven, resulting in seismic shifts in the world table of agricultural production over the past few decades and prospects for continuing shifts over the decades to come.

A half-century ago, today's high-income countries dominated agricultural production and public agricultural research. In the 50 years since then, these countries have shrunk in relative global importance both as agricultural producers and in terms of agricultural research. In counterpoint, the middle-income countries—especially China and Brazil—have grown in importance both as agricultural producers and as performers of agricultural research. These countries have significantly reduced the relative role of agriculture in their own economies while rising to a position of dominance within the global agricultural economy—mirroring the status of today's high-income countries a half century ago. Meanwhile, many of the world's poorest countries continue to lag behind in agricultural production and productivity, in agricultural research, and in making the overall economic transition away from agriculture.

These uneven developments, and the associated systemic seismic shifts in agricultural production, productivity, and spending patterns for agricultural research mean that the world, especially the world's poor, will increasingly depend on the middle-income countries for agricultural innovations and abundance. The shifted shares of the world's agricultural science will have implications over decades to come for the balance of research undertaken, global patterns of productivity and prices, competitiveness and comparative advantage, the mix and quality of food and other agricultural products produced, and the livelihoods of farmers and their families. Even if we see a reversal of research investment trends in the high-income countries, both toward higher levels of agricultural research spending and toward a renewed focus on sustaining and increasing crop yields and other dimensions of agricultural productivity—which does not seem very likely—today's middle-income countries like China, India, and Brazil will increasingly determine the future path of poverty and hunger in the world and the vulnerability of the poor to food price shocks of the kind experienced in 2008 and 2012. These countries are now poised potentially to play a role in the coming 50 years that was played by today's high-income countries in the past 50 years.

■ *We are grateful for research assistance provided by Connie Chan-Kang, Ali Bittinger, and Matt Andersen, and for helpful comments from Alex McCalla, Alan Olmstead, Prabhu Pingali, and the journal editors. The work for this project was partly supported by the University of California; the University of Minnesota; the HarvestChoice initiative, funded by the Bill and Melinda Gates Foundation; and the Giannini Foundation of Agricultural Economics.*

References

Alston, Julian M., Matthew A. Andersen, Jennifer S. James, and Philip G. Pardey. 2010. *Persistence Pays: U.S. Agricultural Productivity Growth and the Benefits from Public R&D Spending.* New York: Springer.

Alston, Julian M., Matthew A. Andersen, Jennifer S. James, and Philip G. Pardey. 2011. "The Economic Returns to U.S. Public Agricultural Research." *American Journal of Agricultural Economics* 93(5): 1257–77.

Alston, Julian M., Bruce A. Babcock, and Philip G. Pardey. 2010. "Shifting Patterns of Global Agricultural Production and Productivity: Synthesis and Conclusion." Chap. 15 in *The Shifting Patterns of Agricultural Production and Productivity Worldwide,* edited by J. M. Alston, B. Babcock, and P. G. Pardey. Ames, IA: Center for Agricultural and Rural Development. http://www.card.iastate.edu/books/shifting_patterns/.

Alston, Julian M., Jason M. Beddow, and Philip G. Pardey. 2009. "Agricultural Research, Productivity, and Food Prices in the Long Run." *Science* 325(5945): 1209–10.

Alston, Julian M., Will J. Martin, and Philip G. Pardey. Forthcoming. "Influences of Agricultural Technology on the Size and Importance of Food

Price Variability." Chapter in *Economics of Food Price Volatility*, edited by J-P. Chavas, D. Hummels, and B. Wright. University of Chicago Press.

Alston, Julian M., and Philip G. Pardey. 1996. *Making Science Pay: The Economics of Agricultural R&D Policy*. Washington, D.C.: American Enterprise Institute Press.

Ball, Eldon Ball, David Schimmelpfenning, and Sun Ling Wang. 2013. "Is U.S. Agricultural Productivity Growth Slowing." *Applied Economic Perspectives and Policy* 35(3): 435–50.

Beintema, Nienke, and Gert-Jan Stads. 2008. "Measuring Agricultural Research Investments: A Revised Global Picture." ASTI Background Note. Washington, D.C.: International Food Policy Research Institute.

Beintema, Nienke, Gert-Jan Stads, Keith Fuglie, and Paul Heisey. 2012. *ASTI Global Assessment of Agricultural R&D Spending*. Washington, D.C.: International Food Policy Research Institute.

Bervejillo, J., Julian M. Alston, and K. P. Tumber. 2012. "The Returns to Public Agricultural Research in Uruguay." *Australian Journal of Agricultural and Resource Economics* 56(4): 475–97.

Butzer, R., Y. Mundlak, and D. F. Larson. 2012. "Measures of Fixed Capital in Agriculture." Chap. 15 in *Productivity Growth in Agriculture: An International Perspective,* edited by K. O. Fuglie, S. L. Wang, and V. E. Ball. Wallingford: CAB International.

Craig, Barbara J., Philip G. Pardey, and Johannes Roseboom. 1997. "International Productivity Patterns: Accounting for Input Quality, Infrastructure, and Research." *American Journal of Agricultural Economics* 79(4): 1064–76.

de Gorter, Harry, and David R. Just. 2009. "The Economics of a Blend Mandate." *American Journal of Agricultural Economics* 91(3): 738–50.

FAOSTAT. 2013. A database of the Food and Agriculture Organization of the United Nations (FAO). Retrieved January 2013 from http://faostat.fao.org.

Fuglie, Keith O. 2012. "Productivity Growth and Technology Capital in the Global Agricultural Economy." Chap. 16 in *Productivity Growth in Agriculture: An International Perspective*. K. O. Fuglie, S. L. Wang, and V. E. Ball, eds. Wallingford: CAB International, 2012.

Fuglie, Keith O., Paul W. Heisey, John L. King, Carl E. Pray, Kelly Day-Rubenstein, David Schimmelpfennig, Sun Ling Wang, and Rupa Karmarkar-Deshmukh. 2011. *Research Investments and Market Structure in the Food Processing, Agricultural Input, and Biofuel Industries Worldwide*. Report No. ERR-130, Economic Research Service, US Department of Agriculture.

Fuglie, Keith O., Sun Ling Wang, and V. Eldon Ball eds. 2012. *Productivity Growth in Agriculture: An International Perspective*. Wallingford: CAB International.

Fuller, Kate B., and Julian M. Alston. 2012. "The Demand for California Wine Grapes." *Journal of Wine Economics* 7(2): 192–212.

Gardner, Bruce L. 1983. "Changing Economic Perspectives on the Farm Problem." *Journal of Economic Literature* 30(1): 62–101.

Gardner, Bruce L. 2002. *American Agriculture in the Twentieth Century: How It Flourished and What it Cost*. Cambridge MA: Harvard University Press.

Grantham, George. 1984. "The Shifting Locus of Agricultural Innovation in Nineteenth-Century Europe: The Case of the Agricultural Experiment Stations." In *Technique, Spirit, and Form in the Making of the Modern Economies: Essays in Honour of W. N. Parker*. Edited by G. Saxonhouse and C. Wright. Greenwich, CT: JAI Press, *Research in Economic History*, Supplement 3, pp. 191–214.

Griliches, Zvi. 1963. "The Sources of Measured Productivity Growth: Agriculture, 1940–1960." *Journal of Political Economy* 71(4): 331–46.

Hayami, Yujiro, and Vernon W. Ruttan. 1971. *Agricultural Development: An International Perspective*. Baltimore: Johns Hopkins University Press. (Reprinted 1985.)

International Fertilizer Association (IFA). 2013. Statistics. Retrieved January 2013 from http://www.fertilizer.org/ifa/HomePage/STATISTICS.

Jorgenson, Dale W., and Zvi Griliches. 1967. "The Explanation of Productivity Change." *Review of Economic Studies* 34(3): 249–83.

Juma, Calestous. 1989. *The Gene Hunters: Biotechnology and Scramble for Seeds*. Princeton University Press.

Kislev, Yoav, and Willis Peterson. 1981. "Induced Innovation and Farm Mechanization." *American Journal of Agricultural Economics* 63(3): 562–65.

Lapan, Harvey, and GianCarlo Moschini. 2012. "Second-best Biofuel Policies and the Welfare Effects of Quantity Mandates and Subsidies." *Journal of Environmental Economics and Management* 63(2): 224–41.

Olmstead, Alan L., and Paul W. Rhode. 2007. "Biological Globalization: The Other Grain Invasion." Chap. 5 in *The New Comparative Economic History: Essays in Honor of Jeffrey G. Williamson*, edited by T. J. Hatton, K. H. O'Rourke, and A. M. Taylor. Cambridge, MA: MIT Press.

Olmstead, Alan L. and Paul W. Rhode. 2009. *Creating Abundance: Biological Innovation and American Agricultural Development*. Cambridge University Press.

Ozgediz, Selcuk. 2012. *The CGIAR at 40: Institutional Evolution of the World's Premier Agricultural Research Network*. Washington, DC: CGIAR Fund Office.

Pardey, Philip G., Julian M. Alston, and Connie Chan-Kang. 2013. *Public Food and Agricultural Research in the United States: The Rise and Decline* *of Public Investments, and Policies for Renewal*. Washington, DC: AGree.

Pardey, Philip G., Julian M. Alston, Jason E. Christian, and Shenggen Fan. 1996. *Hidden Harvest: U.S. Benefits from International Research Aid*. Food Policy Report, Washington, D.C.: International Food Policy Research Institute.

Pardey, Philip G., Julian M. Alston, and Vernon W. Ruttan. 2010. "The Economics of Innovation and Technical Change in Agriculture." Chap. 22 *Handbook of Economics of Technical Change*, edited by B. H. Hall and N. Rosenberg. Amsterdam: Elsevier.

Pardey, Philip G., and Jason M. Beddow. 2013. *Agricultural Innovation: The United States in a Changing Global Reality*. CCGA Report. Chicago: Chicago Council on Global Affairs.

Pardey, Philip G., and Nienke M. Beintema. 2001. "Slow Magic: Agricultural R&D a Century after Mendel." Food Policy Report, International Food Policy Research Institute.

Pardey, Philip G., and Connie Chan-Kang. Forthcoming. *Public and Private R&D for Food and Agriculture in Rich Countries, 1960–2009*. InSTePP Brief. St. Paul: University of Minnesota.

Pardey, Philip G., Connie Chan-Kang, and Steven Dehmer. Forthcoming. *Global Food and Agricultural R&D Spending, 1960–2009*. InSTePP Report. St. Paul: University of Minnesota.

Pardey, Philip G., Bonwoo Koo, Jennifer Drew, Jeffrey Horwich, and Carol Nottenburg. 2013. "The Evolving Landscape of Plant Varietal Rights in the United States, 1930–2008." *Nature Biotechnology* 31(1): 25–29.

Rao, Xudong, Terrance M. Hurley, and Philip G. Pardey. 2012. "Recalibrating the Reported Rates of Returns to Food and Agricultural R&D." Staff Paper. St. Paul: University of Minnesota, Department of Applied Economics. http://ageconsearch.umn.edu/handle/135018.

Ravallion, Martin, Shaohua Chen, and Prem Sangraula. 2007. "New Evidence on the Urbanization of Global Poverty." *Population and Development Review* 33(4): 667–701.

Ruttan, Vernon W. 1982. *Agricultural Research Policy*. Minneapolis: University of Minnesota Press.

Ruttan, Vernon W. 2002. "Productivity Growth in World Agriculture: Sources and Constraints." *Journal of Economic Perspectives* 16(4): 161–84.

Ryerson, Knowles A. 1933. "History and Significance of the Foreign Plant Introduction Work of the United States Department of Agriculture." *Agricultural History* 7(3): 110–28.

Schultz, Theodore W. 1956. "Reflections on Agricultural Production Output and Supply." *Journal of Farm Economics* 38(3): 748–62.

Swinnen, Johan F. M., Kristine Van Herck, and Liesbet Vranken. 2010. "Shifting Patterns of Agricultural Production and Productivity in the Former Soviet Union and Central and Eastern Europe." Chap. 10 in *The Shifting Patterns of Agricultural Production and Productivity Worldwide*, edited by J. M. Alston, B. A. Babcock, and P. G. Pardey. CARD-MATRIC Electronic Book. Ames, IA: Center for Agricultural and Rural Development. http://www.matric.iastate.edu/shifting_patterns/.

Timmer, C. Peter. 2009. *A World without Agriculture: The Structural Transformation in Historical Perspective*. Washington, D.C.: American Enterprise Institute Press.

US Department of Agriculture (USDA), Economic Research Service (ERS). 2012. "Food Expenditure Series." Washington, D.C.: United States Department of Agriculture. Retrieved from http://www.ers.usda.gov/data-products/food-expenditures.aspx.

US Department of Agriculture, National Agricultural Statistics Service (USDA, NASS). 2013. Data and Statistics - Quick Stats database. Retrieved from http://www.nass.usda.gov/Data_and_Statistics/index.asp. Accessed April 2013.

World Bank. 2007. *World Development Report 2008: Agriculture for Development*. Washington, DC: World Bank.

World Bank. 2011. *World Development Report 2011: Conflict, Security, and Development*. Washington, DC: World Bank.

World Bank. 2012. *World Development Indicators Online*. Washington, D.C.: World Bank Retrieved December 2012 from http://databank.worldbank.org/ddp/home.do?Step=12&id=4&CNO=2.

World Bank. 2013. "Data-Methodologies." Washington, DC: World Bank. Available at http://data.worldbank.org/about/data-overview/methodologies.

Wright, Brian D. 2012. "Grand Missions of Agricultural Innovation." *Research Policy* 41(10): 1716–28.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 147–166

American Farms Keep Growing: Size, Productivity, and Policy

Daniel A. Sumner

According to the official US government definition, about two million farms operate in the United States, a total that has changed little for more than three decades. However, the US government defines a "farm" as an operation that produced or normally would produce at least $1,000 in gross value of output per year (O'Donoghue, Hoppe, Banker, and Korb 2009). The tiny threshold of $1,000 in sales represents an agricultural output of less than two acres of corn, less than one-half of a milk cow, and less than half of one litter from one mother sow. This inclusive threshold has public relations and political rationales. Farm lobby groups, political officials at the US Department of Agriculture, and elected representatives laud the efforts of "two million American farms" to feed the world or support the rural economy.

In fact, most of these "farms" contribute approximately nothing or less than nothing to these objectives. About 600,000 farms, about 30 percent of the two million, have any significant farm production, and about 120,000 farms, or 6 percent of the total, produce three-quarters of all US farm output. Using the $1,000 threshold to define a farm, about half of all farms in the United States gross less than $5,000 per farm. As a group, the tiny farms have expenses well in excess of revenue. Based on survey data from the National Agricultural Statistics Service of the US Department of Agriculture (USDA) and the Census of Agriculture, operators of the tiny farms garner their livelihood from other sources including employment in the nonfarm economy and retirement income. Indeed, one reason the average age of farmers

■ *Daniel A. Sumner is the Frank H. Buck, Jr. Professor of Agricultural and Resource Economics, University of California at Davis, Davis, California. His email address is dasumner@ucdavis.edu.*

http://dx.doi.org/10.1257/jep.28.1.147 doi=10.1257/jep.28.1.147

has been over 55 for decades is that many individuals continue to operate very small farms as a retirement activity after they have left nonfarm employment. Despite negative returns from farming, operators of tiny farms have incomes well above the national average—for them, in most years, farming is a costly avocation. In 2012, a year with record net farm income, farms with sales less than $100,000 generated negative aggregate net farm income of about –$2,500 per farm (Economic Research Service, USDA, 2013).

In this article, I concentrate on commercial farms—those that typically provide some positive net income and might engage an operator on a full-time basis. The size of these farms has more than doubled over the past two decades. Acres of corn per corn farm rose to 600 acres in 2007 from 200 acres in 1987 for the farm at the center of the production distribution, meaning half of production came from larger farms and half from smaller farms. Corn yield per acre rose by more than one-third over the same period. Acres per farm at the center of the production distribution for cotton, rice, soybean, and wheat also more than doubled. For some commodity industries, farm size growth has transformed the industry. For example, the hog operation at the center of the production distribution sold 30,000 hogs per year in 2007 compared to 1,200 in 1987. The central-sized dairy herd size was 570 in 2007 compared to 80 in 1987. Since aggregate acreage, inflation-adjusted farm revenues, and farm value added have changed relatively little in recent decades, expansion of farm size and farm consolidation become synonymous.

This article summarizes the economics of commercial agriculture in the United States, focusing on how growth in farm size and other changes in size distribution have changed in recent decades. I also consider the relationships between farm size distributions and farm productivity growth and farm subsidy policy. The search for causation among these relationships remains challenging. To focus on the recent evolution of farming in the United States, our time horizon is roughly three decades, from the early 1980s to the present. For a sweeping view of developments in US agriculture during the twentieth century, Gardner (2002) is a useful starting point. Economists at the Economic Research Service of the US Department of Agriculture have advanced our knowledge of farm size and the industrial organization of farming along with the relationships between farms and farm supply and marketing firms. They have used unique USDA data, including individual farm records from successive rounds of the Census of Agriculture and the annual Agricultural Resource Management Survey (ARMS). This article draws upon and cites much of this work.

The Size Distribution of American Farms

The growth of farm size is not a simple proposition to document with aggregate data. Farm size is often measured by sales or by quantities of output or land and other inputs, but none of these measures suits all questions. Farm sales data are affected directly by changes in relative prices over time, and commodities differ in value added and also in management effort per unit of gross sales. Land area can

Table 1
The Size Distribution of Farms by Gross Sales

	Number and share of farms (in 1,000s and percent)			*Aggregate sales and shares of aggregate sales for firms in the sales category (in 1,000s of dollars and percent)*		
Sales category	*1987*	*1997*	*2007*	*1987*	*1997*	*2007*
< $25k	1,099.7 (54.7%)	1,017.7 (53.2%)	1,355.9 (61.5%)	8,222.6 (6.1%)	6,943.2 (3.5%)	6,314.8 (2.1%)
$25k–$50k	219.6 (10.5%)	170.7 (8.9%)	154.7 (7.0%)	7,868.7 (5.8%)	6,084.3 (3.1%)	5,480.5 (1.8%)
$50k–$100k	218.1 (10.4%)	158.2 (8.3%)	125.5 (5.7%)	15,661.4 (11.5%)	11,346.5 (5.8%)	8,961.3 (3.0%)
$100k–$250k	202.6 (9.7%)	189.4 (9.9%)	147.5 (6.7%)	31,178.2 (22.9%)	30,143.3 (15.3%)	24,212.9 (8.1%)
$250k–$500k	61.1 (2.9%)	87.8 (4.6%)	93.4 (4.2%)	20,739.7 (15.2%)	30,505.2 (15.5%)	33,409.9 (11.2%)
$500k–$1 million	20.9 (1.0%)	42.9 (2.2%)	60.8 (2.8%)	14,076.3 (10.3%)	29,365.1 (14.9%)	42,690.8 (14.4%)
$1 million–$2.5 million*	11.1 (0.5%)	19.1 (1.0%)	40.4 (1.8%)	37,876.0 (27.8%)	27,938.7 (14.2%)	60,549.3 (20.4%)
$2.5 million–$5 million*	N/A	4.1 (0.2%)	9.6 (0.4%)	N/A	13,802.4 (7.0%)	32,299.5 (10.9%)
> $5 million*	N/A	2.8 (0.1%)	5.5 (0.3%)	N/A	40,369.9 (20.5%)	82,951.0 (27.9%)

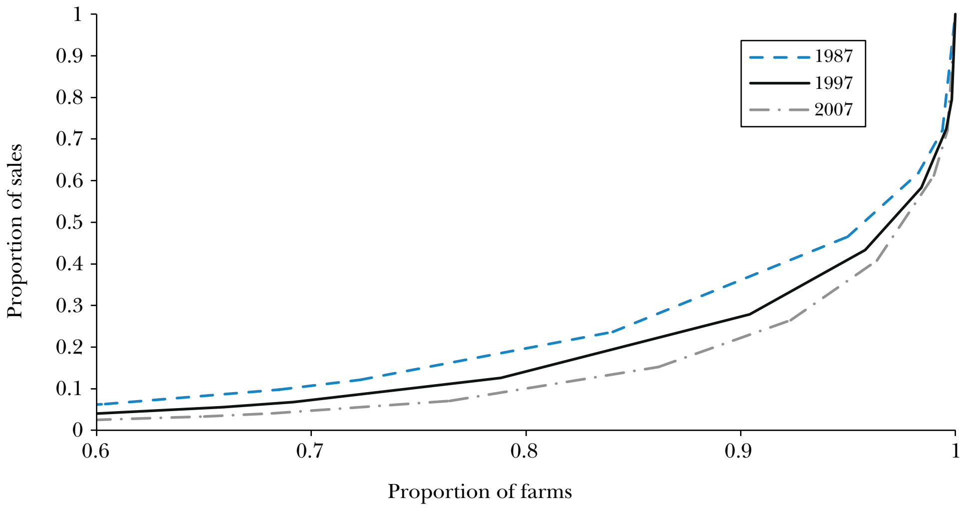
Source: United States Department of Agriculture, Census of Agriculture, 2007 Table 2: "Market Value of Agricultural Products Sold Including Landlord's Share and Direct Sales 2007 and 2002," 1997 Table 2: "Market Value of Agricultural Products Sold and Direct Sales: 1997, 1992 and 1982."
*($1MM+ for 1987), 1987 Census of Agriculture did not provide the numbers of farms of sales for separate categories over $1 million.

work for a particular commodity (or commodity mix) and location. Within specific commodity industries and regions, farm size is typically measured by input quantities (like harvested cropland or number of milk cows), output quantities (bushels of corn or hundredweight of milk), or value of output. In theory, farm size might also be measured by value added, net revenue, or total returns to management and fixed assets, but such data are not widely available. Farm size is almost never measured by the value of the capital stock or the number of employees, the dimensions that are commonly used in measuring size across firms outside of agriculture.

By any measure, commercial US farm sizes have risen dramatically in the last few decades. This growth in size applies across the distribution of farm sizes: midsized commercial farms have gotten larger, and so have larger farms. As commercial farms grow larger, the number of tiny farms has been increasing slightly, so mean and median farm sizes for official aggregate data have been getting larger only slowly.

Table 1 presents aggregate data on the share of farms across sales classes for census years 1987, 1997, and 2007. The left three columns show the number and

Figure 1
Changing Distribution of Farm Sizes, 1987, 1997, and 2007
(farm size measured in terms of sales)



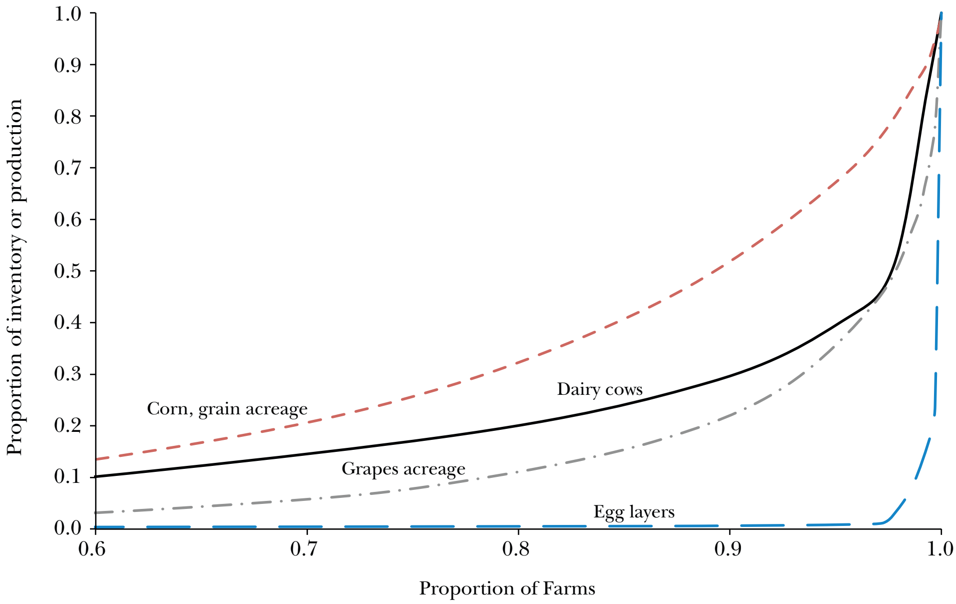
Source: US Department of Agriculture, Census of Agriculture, 1987, 1997, and 2007, http://www.agcensus.usda.gov/index.php.

share of farms in each category of sales; the right three columns show aggregate sales and shares of aggregate sales from firms in those sales classes. Farms that gross less than $25,000 per farm, making up approximately 60 percent of all farms, only accounted for about 2 percent of US farm output in 2007. (Farm output prices were stagnant from 1982 to 2002, but rose by about 40 percent from 2002 to 2007, and have risen another third since.) At the upper end of the distribution, the share of total sales from farms with sales of $1 million or more rose from less than 30 percent to 60 percent in 2007. Table 1 presents categories in nominal dollars. O'Donoghue et al. (2011) adjusted sales categories to be in constant 2007 dollars and found that farms producing more than $1 million of output (in 2007 constant dollars) rose by 243 percent from 1982 to 2007 to reach 55,500 farms.

Figure 1 shows the shifting concentration over time. At the bottom left, the smallest 60 percent of farms by sales generate about 5 percent of total sales over time. The smallest 60 percent of farms as well as 70 percent, 80 percent, and 90 percent, all have a lower share of total sales in 1997 than in 1987 and a lower share in 2007 than in 1997.

Measuring farm size by inputs, like cropland, requires care concerning differences across crops. As an extreme example, operating 1,000 acres of strawberries would entail many employees and tens of millions of dollars in revenue, whereas 1,000 acres of wheat entails part-time employment for a single operator and revenue of less than $0.5 million. Corn and grape farms are often measured by acreage.

Figure 2
Distribution of Farm Size for Corn, Grapes, Dairy, and Egg Industries
(farm size measured in acreage for corn and grapes; in livestock numbers for dairy cows and egg layers)



Source: USDA, Census of Agriculture, 2007, http://www.agcensus.usda.gov/index.php.

Dairy and egg farms are typically measured by number of livestock. Figure 2 shows how the size distributions of these types of farms differ. The corn industry is least concentrated, with the largest 20 percent of farms operating on about 70 percent of the acreage. About 10 percent of the farms operate on about half the corn acreage.

Among dairy farms, 20 percent of the farms milk about 80 percent of the total cows, and less than 5 percent of the farms milk more than half of the total cows. Small commercial dairy farms milk fewer than 100 cows and have milk sales gross revenue of perhaps $300,000. Larger dairy farms have in the range of 10,000 cows and milk sales in the range of $30 million. Dairy farms differ by more than just herd size, and degree of vertical integration matters. For example, some grow their own feed or breed replacement heifers, but herd size remains a useful measure of overall size (Sumner and Wolf 2002).

About 90 percent of grape farm acreage is on the largest 20 percent of grape operations, and 80 percent of acreage is on the largest 10 percent. The egg farm size distribution is clearly the most concentrated shown in Figure 2. The largest egg farms tend to have operations in multiple locations and millions of hens. But even in the egg business, where markets are national and regional, dozens of farms compete in most markets.

Farm consolidation raises few market power issues. Farms that produce the bulk of US output still number in the tens of thousands, and few are large enterprises by nonfarm standards. Some farms do have sales in the tens of million dollars. These tend to be in industries with low value added per unit of sales, such as livestock feeding, where cost of young livestock and feed comprise up to 90 percent of gross sales. The crop industries with high-sales farms include fruits, tree nuts, and vegetables. Even in these industries, dozens or even hundreds of farms compete in the relevant markets.

Are the large farms generally operated as parts of large publicly held corporations that primarily engage in nonfarm enterprises? The available data say no. In 2007, 1 percent of farms with sales of at least $100,000 were organized as nonfamily-held corporations, while 71 percent are held by individuals (according to the Census of Agriculture). Family farms, where the operator or immediate family members own at least half the farm, account for about 87 percent of crop output. Of nonfamily farms, corporations produce about half of the value of crop output. Corporate operation of livestock feeding accounts for a much larger share of output, probably exceeding 50 percent.

Market power in agriculture is typically exercised by downstream firms or is attributable to consumer product brands that may be controlled by large farms or groups of farms operating as marketing firms. Some farm-based market power in American agriculture may also derive from location-specific attributes affecting consumer willingness to pay, such as the case of winegrapes from a particular vineyard, but these are rare cases, similar to the market power of a few extremely talented athletes or celebrities.

At the same time that the skewness of farm size distributions has increased, shares of output and numbers of farms have been declining in size categories that were considered mid-size two decades ago. Does that mean farm size distributions are becoming bimodal, with some huge farms, many tiny farms, and few in-between? In fact, there is little or no evidence of multiple modes in farm size distributions. Whole distributions have shifted to the right such that formerly mid-sized commercial farms are now small and what would have formerly been a large farm is now considered mid-sized. In the only study using appropriate nonparametric tools (Wolf and Sumner 2001), we rejected the bimodal hypothesis for US dairy farm data, a farm commodity industry where measurement of size is among the least complicated.

What about the small organic farms that market directly to consumers through farmers' markets, home delivery, or to local retailers—the farms that seem to dominate urban elite consciousness? Though such farms may frequently appear in the popular press, organic production amounts to 1–2 percent of farm output, mainly concentrated in milk and fresh produce and on the East and West Coasts. Moreover, large farms that ship products nationally have a large share of the organic market. Farm output from *small farms* that *supply local markets* with *organic products* comprises the intersection of three already small sets and is simply not a significant part of agricultural production in the United States.

Finally, we may consider farm size distributions in relation to distributions of other business firms. In general, even large farms are small businesses. Indeed, most business firms are also small, many employing no workers in addition to the operator and most with very few employees (Evans and Leighton 1989; Cabral and Mata 2003; Luttmer 2011). The US Small Business Administration (undated) defines a "small" business in the context of the industry. In manufacturing, the maximum number of employees for a "small" firm may range from 500 to 1,500, depending on the type of product manufactured. In wholesaling, the maximum number of employees for a "small" firm may range from 100 to 500, depending on the particular product being provided. In services, the annual receipts of a "small" firm may not exceed $2.5 to $21.5 million, depending on the particular service being provided. In general construction, annual receipts for a "small" firm may not exceed $13.5 to $17 million.

In farming, the Small Business Administration criteria for annual receipts of a "small" farm may not exceed $0.5 to $9.0 million, depending on the agricultural product. For certain farm product industries, such as grains and oilseeds, the limit for a small business is set much below those in other industries. But the sales limit for "small" for livestock feeding is set at a range comparable to services or construction. Manufacturing is categorized for size by number of employees, and the minimum number of workers for a small business is well above that of all but a tiny share of commercial farms.

The most obvious difference between farming and manufacturing or services is that, unlike other industries, there are no farms with tens of thousands of employees that operate on a global scale. No firms that are primarily engaged in farming are among the 500 largest firms in the economy, and very few farms have familiar company names.

What Drives Differences in Farm Size and Shifts in Farm Size Distributions Over Time?

Commercial farms have increased in size by every measure, both for US farming as a whole and across the full range of commodities. Characterizing and explaining farm size distributions and their change over time is complex, and economists have not yet developed fully satisfying models to account for observed patterns. This section considers the attempts to understand farm size distributions.

Among the first questions is why farm size distributions differ from what is typical in the rest of the economy. In particular, why are there few very large farms? One may suggest hypotheses to explain this fact, but no studies have provided satisfying empirical tests. The most promising avenues of explanation involve cost of monitoring and rewarding performance in large farms and gains to specific local knowledge. For example, it may be more difficult to monitor effort and ability when weather and biological variations dominate performance, especially with few workers per location. It may also be more difficult to integrate production and marketing when generic products dominate and the return to branding is small. Animal feeding establishments operated by firms such as Cargill and Smithfield are examples where local conditions are less important and standardization of performance expectations allow a large operation to match efficiencies of independent farms. Also, although most large wineries buy most of the grapes they process, some large firms with well-known wine brands produce grapes on winery-operated farms.

There has been relatively little cross-fertilization between the general literature on firm size and the farm size literature. Hall (1987) pointed out how farm size distributions had much in common with other industries. At that time, Gibrat's law, which holds that the size of a firm and its growth rate are independent, captured the size distribution of firms and farms fairly well, especially for smaller firms. While economists who study farm size refer to the early models developed for size and growth of business firms (such as Lucas 1978; Jovanovic 1982; Evan and Jovanovic 1989), they generally have not tested such models with data on farms or compared their results to those for business firms (for example, Evans 1987; Cabral and Mata 2003; Angelini and Generale 2008).

In one of the most prominent contributions to the literature on farm size, Kislev and Peterson (1982) focused on land operated per farm as the measure of farm size. They assumed, consistent with the facts across much of the US Midwest during their period of investigation, that each farm had a single operator who also supplied most of the labor on the farm. Using a series of plausible simplifying assumptions, they showed how an increase in the income potential in nonfarm occupations would raise "machinery-capacity" per farm (because machine capital would be substituted for labor). Then, as the capacity of machinery to cultivate, plant, or harvest more acreage in limited time-windows increased, acreage per farm and per farmer would increase. The result was that the opportunity cost of earning labor off the farm would generate more cropland per farm operator. This model also implied fewer farms, given a fixed or almost fixed amount of cropland. Kislev and Peterson (1982) showed that data from US agriculture in the middle of the twentieth century was consistent with their assumptions about substitution between land, labor, and capital, and their model could help explain the pattern of crop farm growth in the United States. No published work has tested the relevance of nonfarm wages for farm size using the more recent data.

Most commercial farms in the United States no longer match the stylized facts that Kislev and Peterson (1982) used in generating their result. For example, commercial farms employ more than one tractor and harvester (or buy equipment services) and operate fields at several locations, and farm managers often do not operate farm equipment themselves. Many farms have several manager/operators who share returns. Real wages for hired farm workers (and other workers) have stagnated, while many farm tasks have increased in complexity and technical expertise required. At the same time, the human capital of farmers and their opportunity cost of forgoing off-farm careers have risen. That said, many farmers are middle-aged or older and have little nonfarm work experience, so their employment opportunities at nonfarm occupations are now limited. Also, with the amount of labor and land per farm variable rather than fixed, labor-saving technology may increase land-to-hired-labor ratios without affecting the total acreage, production, or revenue per farm.

Determinants of farm size in current farming may be derived more from managerial ability and selection for better managers, as in models of Lucas (1978) and Jovanovic (1982). The following nine stylized facts and hypotheses are useful in summarizing ideas about how human capital and managerial capabilities affect the economics behind observed patterns of commercial farm size and growth in the United States. They may serve as building blocks for those seeking to model these patterns.

First, farm average cost curves are generally L-shaped. Costs fall rapidly as size expands for a limited range, but after a minimum size, costs decline very gradually, if at all, and farms are distributed across a range of sizes based on some criteria other than economies and diseconomies of scale. Without further elaboration, such a model is silent about why firms larger than the minimum grow or why farms have different sizes. Although it is difficult to measure the opportunity cost of the farm owner/operator in terms of earnings at an alternative occupation, it seems likely that farms differ in optimal size, and somewhat smaller farms may remain competitive with larger farms that have slightly lower measured costs, because the smaller farms have a lower opportunity cost of using their human capital in farming.

This pattern of L-shaped average cost curves suggests the presence of technology that is not scale neutral, perhaps because of lumpy physical assets, such as tractors or planting machinery, milking parlors, or hen housing facilities, or lumpy managerial capital for firms with one or very few managers. However, the evidence suggests that these economies of scale seem to almost level out at some medium level of input. The evidence is mixed. For example, using a cross section of dairy farms in 2000, Mosheim and Lovell (2009) find scale economies throughout the range of their data. While, using individual records from several rounds of the Census of Agriculture, Melhim, O'Donoghue, and Shumway (2009) find that scale economies are relatively unimportant at larger farm sizes. They also find greater diversification by large firms, which suggests economies of scope—and managerial skill may well be useful in discovering and implementing such economies of scope. Recent accounting data on a larger random sample of farms shows rates of return on equity rising through all farm acreage classes for corn, soybeans, wheat, fruits and nuts, and vegetable and melon farms (MacDonald, Korb, and Hoppe 2013, table 5). Larger crop farms use less labor and less capital per acre.

Second, although the cost curves are L-shaped, changes in technology have meant that the minimum-cost farm size has been shifting out over time. This observation is consistent with growth of farm size, and with entrants to the farm sector having larger size at time of entry. There is evidence of this occurring for both crop and livestock industries (MacDonald, Korb, and Hoppe 2013).

Third, consistent with recognition of economies of scope, farms also grow by adding commodity enterprises with different specific patterns of labor, management, and specialized capital across the year and across the farm. Corn and soybeans are grown in rotation on the same land to increase yields, which means that farms often grow both each year. Reasons not to specialize in a single crop each year include different seasonal patterns that allow more complete annual use of machinery and management. Moreover, even similar land in the same region is often best suited for particular crops, so as farms expand they naturally occupy land that may be suited for different commodities. Of course, diversification across enterprises also contributes to risk management, especially for vertically integrated commodities such as feed and livestock production.

Fourth, larger farms typically rent land, while smaller farms own more of the land upon which they operate (based on Census of Agriculture data; see also MacDonald, Korb, and Hoppe 2013). Larger farms tend to use their capital for operation and rely on outsiders—often former farmers and farm heirs—to invest in farmland ownership. This separation of farm operation from land ownership involves a variety of contracts, which range from hired managers earning salaries and incentive bonuses, through a variety of share tenancy arrangements (where tenants and landlords share costs and returns), to cash rental contracts where the landlord is involved in few of the farm decisions and faces none of the variability of farm costs and returns. The contract arrangements differ by the particular conditions and capabilities of landlords and tenants (Allen and Lueck 2002).

Fifth, managerial capability in agriculture may be a lumpy input, which in turn limits scale economies and the potential for farm growth (Tolley 1970; Alvarez and Arias 2003). In Sumner and Leiby (1987), we showed that human capital of operators determined farm size and improving managerial ability among dairy farmers helped explain increasing average herd size and growth. Hierarchical management systems may be difficult to employ in farming, because so much of the production effort is highly specific to the location in which it happens. This restriction on exercise of managerial capability, and a reliance on family and ownership relationships, limits farm size compared to the very large firms in other industries. Where standardization and monitoring allow operations in multiple locations and with many managers, such as in the livestock feeding and egg industries, farms can and do become much larger.

Sixth, when the payoff to nonfarm opportunities for farmers rises relative to the cost of capital, the result is more capital per farmer. As real interest rates have fallen over the past 30 years and returns to human capital have risen, capital per farm rises as well as capital per unit of hired labor (which has received largely stagnant wages). When capital substitutes for farm operator and manager time rather than routine hired labor, farmland operated, gross sales, and value added per farmer and per farm all rise. This insight is related to the Kislev and Peterson (1982) hypothesis, but focuses on farmer management, not farm labor.

Seventh, managerial capability has become more similar between farming and nonfarm occupations, especially in organizations with specialized technical personnel or contracted specialist services, such as from animal nutritionists or pest consultants. Thus, compared to technical farming skills, managerial capability is more likely to be human capital that is general across occupations and less farm-industry-specific.

The implication is that farms must be large enough to attract and retain farm operators and managers who have high opportunity costs off the farm. Moreover, higher returns to management at larger farms create incentives for farm consolidation.

Eighth, farms with more-capable managers are larger for two reasons: a) because farmers who are more capable compete effectively for land and other resources and acquire farms previously operated by less-capable farmers; and b) because the returns to managerial capability are high in nonfarm occupations, only those with opportunity to operate large farms, which have scope for relatively high incomes, select farming as an occupation. Moreover, as farm operators with less managerial capability have been replaced by those with higher capability, farm size has grown as a consequence. Without a large operation over which to spread their managerial capabilities, better managers would not enter or continue in farming. These empirical relationships are not axiomatic. In some parts of the world where farm opportunities have been restricted or farming has been otherwise unattractive, the best potential farmers leave the rural areas, and those remaining tend to have below-average capability. Opportunities off the farm draw away the best, and farm size and productivity is left to flounder.

Recent data on farm size, schooling, and managerial capability is compelling. The best recent information comes from analysis of farmers who also work off the farm (Brown and Weber 2013). Of those farms with some off-farm employment, 44 percent of operators of farms with at least $250,000 were employed in managerial and professional occupations compared to 35 percent of operators of smaller farms. About half the operators of the larger farms work off the farm in agriculture or government. In 2010, 58 percent of these employed operators of farms with at least $250,000 in sales (and two thirds of their spouses) had some college, compared to 45 percent of operators of farms with $50,000 to $249,999 in sales.

Technical change and management information systems extend the payoff to managerial capability and the result is higher returns to management in larger farms. The payoffs to finding lower input prices, searching for higher output prices, managing production and price risk, and managing improved productivity of hired labor are all higher for larger operations. Thus, the process of managerial replacement (better managers for worse) complements and is complemented by innovations in technology and information systems that pay dividends to better management.

Finally, ninth, the interaction between technology, managerial replacement, and larger farm size affects the pace of productivity gains just as productivity changes affect the distribution of farm size. Farm size may also be affected by government subsidy policy and regulatory policy. These are the topics of the next two sections.

Farm Size and Productivity

In much of the economy, economists use "productivity" as shorthand for labor productivity (as when they refer to output per worker) and "firm size" to refer to number of employees. For crops, output per acre ("crop yield") is the common

Table 2
National Single Factor Productivity Growth for Grapes, Corn, and Milk

	Average yield		*Average revenue (1,000s of dollars)*		*Compound growth rate (%)*	
Commodity	*1991–93*	*2009–11*	*1991–93*	*2009–11*	*Yield*	*Revenue*
Table grapes/Acre	8.2 tons	11.5 tons	3.7	6.4	1.7	2.7
Wine grapes/Acre	7.4 tons	7.2 tons	2.7	4.4	–0.1	2.4
Corn/Acre	114 bu.	155 bu.	0.26	0.76	1.5	5.5
Milk/Cow	15,441 lbs.	21,022 lbs.	1.9	3.5	1.6	2.9

Source: National Agricultural Statistics Service, US Department of Agriculture.

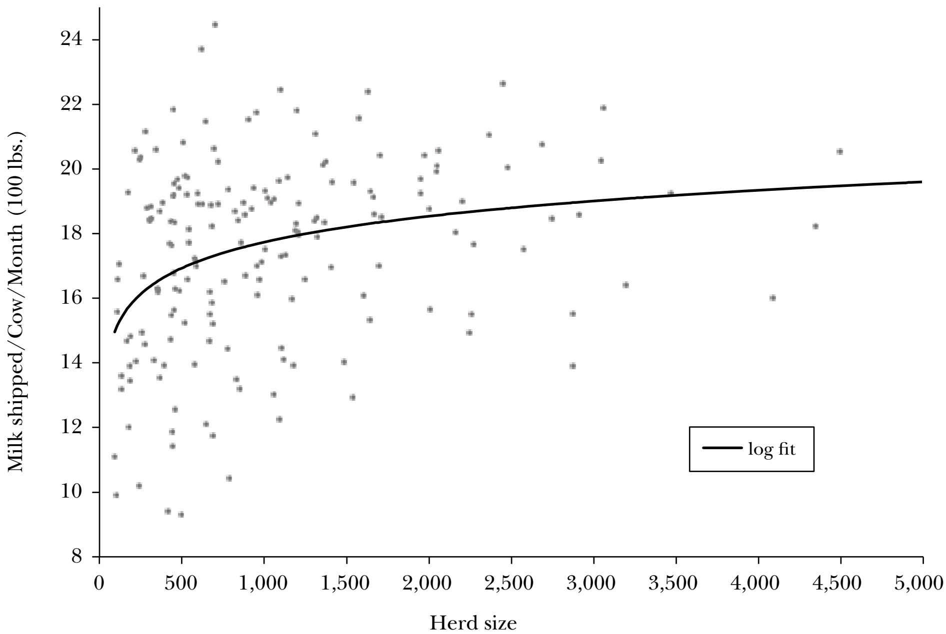
measure of single factor productivity, whereas for livestock, output per animal is common. For example, on US dairy farms, milk per cow is the most common benchmark. For farming as a whole, multifactor or total factor indicators must be used to compare productivity.

Productivity growth has characterized US agriculture for many decades. Rapid increases in crop outputs per acre or outputs per animal apply across almost all commodity groups. Table 2 shows rates of growth over a two-decade period in output and revenue per acre for wine grapes, table grapes, and corn, as well as output and revenue per cow for milk. This group of commodities illustrates the diversity and complexity of single factor productivity measures.

Average milk per cow and corn per acre both rose by about one-third and table grape yield rose by 28 percent in the almost two decades from 1991–1993 to 2009–2011. Wine grape yields, however, fell marginally over this period. Of course, single factor measures have limitations. In the case of wine grapes, changes in product characteristics (and location) mask yield growth for many varieties in many regions. In the case of milk and corn, some yield growth was caused by more feed per cow and more fertilizer per acre that represent movements along isoquants rather than shifts in the isoquant over time.

Applying sophisticated measures of productivity like total factor productivity requires great attention to detail. For example, challenges have included creating indexes of labor and management to account for quality, where much of the labor is entrepreneurial and part time, and operator and family hours of work on farms are not well measured. Aggregating capital across everything from milk cows to tractors and land improvements raises concerns about input quality and quantity. Creating indexes of output to allow aggregating a ton of Cabernet Sauvignon grapes from Napa ($8,000 per ton) with grapes from Fresno ($300 per ton) is equally complex—leaving aside aggregating these with a ton of wheat from Kansas. Notwithstanding these challenges, several teams of researchers have shown that total factor productivity has grown rapidly for US farming and for most significant commodity industries and regions (Alston, Anderson, James, and Pardey 2010; Fuglie, MacDonald, and Ball 2007).

Figure 3
Milk Shipped per Cow, per Month, by Herd Size



Source: California Department of Food and Agriculture, Cost of Production Feedback data, http://cdfa.ca.gov/dairy/uploader/postings/copfeedback/.

The relationships between productivity growth and farm size appear to run through several channels, and in both directions. A large literature focuses mostly on very small farms in poor countries and often finds that crop yields per acre are higher on farms operating less land (Feder 1985; Eastwood, Lipton, and Newell 2010). Higher labor and management inputs per unit of land imply higher output-to-land ratios. On very small farms, the farmer knows his land very well and applies considerable attention to increase yield per acre (Eastwood, Lipton, and Newell 2010).

This pattern is not standard for rich-country agriculture and does not seem to hold in the United States. If the best managers operate larger farms with the best technology, they are likely to have higher productivity. For example, in Iowa corn yields in 2007 increase with acres of corn per farm up to about 1,000 acres. For dairy farms in the San Joaquin Valley of California, the highest milk production region in the United States, there is a clear positive relationship between herd size and milk per cow, as shown in Figure 3. Doubling cows per herd from 500 to 1,000 increases milk production per cow by about 6 percent. Nationally, the gains of larger dairy farms are much more from reduced labor and capital per unit of milk production rather than more milk per cow (MacDonald and McBride 2009).

When larger commercial farms have higher productivity, as farms consolidate and the less-productive operations leave the industry, productivity grows over time. That means part of the measured productivity growth may be due to increases in farm size. The growth in the size of commercial farms, and consolidation that facilitated improved management of farms, contributed to the rapid productivity gains in the half century after 1945. It is not clear if productivity growth associated with these changes has continued to be as rapid, or if more such productivity gains are available (MacDonald, Korb, and Hoppe 2013). Causation in the other direction can also arise from any size bias of improved technology or new technology that favors higher-capability managers who operate larger farms.

Total factor productivity growth in farming continues, but the rate of productivity growth seems to have slowed down since around 1990 (as Alston and Pardey explain in this symposium). The rate of investment in agricultural research and development spending seems to have slowed as well. Other contributors to a growth slowdown may be unmeasured improvements in output quality and more resources devoted to environmental concerns that tend to reduce productivity growth as conventionally measured. Economists have paid considerable attention to these issues, but including them in input and output indexes for productivity growth is inherently difficult.

Because of consolidation in livestock feeding, much of total farm production now comes from farms that seem to have exhausted available scale economies. For example, McBride and Key (2013) document very rapid consolidation and productivity growth that seems to have now exhausted scale economies in hog production. For dairy and field crop operations, rapid consolidation over the past 15 years has corresponded with the period of the productivity growth slowdown. Further research needs to examine productivity growth and farm consolidation patterns across commodity industries and locations in the United States to attempt to assess causation in that relationship.

Farm Subsidies and Farm Size

The evolution of farming in the United States is inevitably connected to the evolution of farm subsidies (Dimitri, Effland, and Conklin 2005; Gardner 1992; Gardner 2002; Sumner, Alston, and Glauber 2010). As commercial farms have become larger and more productive and farm operators have become wealthy relative to other Americans, farm policy in the United States has also evolved. Prominent farm subsidies mostly transfer income to owners of resources on farms that grow the politically favored crops and livestock commodities. Our question here is the extent which farm subsidy policies have influenced farm size distributions overall and for the most heavily subsidized commodities. This has been a topic of vigorous speculation and some research for many decades (Sumner 1985). For example, on one side Key and Roberts (Key and Roberts 2007; Roberts and Key 2008) argue that farm policy has tended to hasten the movement toward larger farms. My own reading of the evidence is that although farm policy has for some commodities leaned against the prevailing winds of consolidation, it has had little overall effect on limiting consolidation of farming—a conclusion that is consistent with evidence assessed by MacDonald, Korb, and Hoppe (2013) and White and Hoppe (2012). After presuming that an objective of farm subsidies was to slow farm consolidation, Eastwood, Lipton, and Newell (2010) came to a similar conclusion for high-income countries overall. They summarize (p. 3374): "In fact, OECD farm support has not overcome the tendency of farm size to grow and of farm numbers and employment to decline." We will discuss the evidence behind these conclusions, after quickly summarizing farm subsidies in the United States.

Farm subsidies differ widely by commodity, but most of the subsidy and the highest subsidy rates still go to what are called the "program crops": grains, cotton, and oilseeds (through government payments), sugar (through trade barriers), and milk (through an array of payments and price regulations). Trade barriers, such as those for frozen concentrated orange juice and fresh market tomatoes, have provided protection to certain commodities for many years. However, most fruits, tree nuts, vegetables, and hay are much less subsidized, and eggs and meat products are largely unsubsidized.

In the old price support programs, the federal government set minimum prices and acquired any supplies not bought at those prices. About three decades ago, these direct price supports began to be replaced by payments that made up the difference between the government-set target and the market price. These payments effectively set a minimum price for farm suppliers, but the market price cleared the market. In 1996, under a program that only economists appreciated, the Congress supplemented the payments tied to specific commodity production and prices with payments based on a farm's history of production of the subsidized crops. Such approximately lump-sum payments reduced production distortions inherent in payments tied to specific commodities. Over the past decade, government agricultural policy has shifted toward heavily subsidized crop insurance (Glauber 2013; White and Hoppe 2012). The Risk Management Agency at the US Department of Agriculture (2013) estimates that government costs were about $15.8 billion for crop year 2012, up from about $9.4 billion for crop year 2011. Insurance programs suffer from substantial moral hazard and adverse selection because farm differences and farm behavior are expensive to monitor. Subsidized crop insurance programs transfer benefits to participating farms, while also offsetting unanticipated variability in prices and yields.

One would expect subsidies to affect which crops are grown and the aggregate supply of farm output, but given the complexity of farm programs, the supply response has not been easy to assess. Aggregate supply functions tend to be inelastic. In the medium-run, even supply response of major program crops such as corn and soybeans remain inelastic because suitable land for these crops is already engaged and because profitable crop rotations planned several years in advance limit adjustments (Hendricks and Sumner, forthcoming). However, for crops that use a small share of potentially suitable land and receive high subsidy rates, such as cotton, medium-run supply functions are more elastic.

The effect of subsidies (including crop insurance) on commercial farm size distributions is harder to assess than the effect of subsidies on aggregate supply response. First, gross empirical relationships do little to illuminate effects of subsidy on size. Major commodity subsidy programs are tailored mostly to field crops that tend to be smaller (by gross sales) than farms producing nonsubsidized commodities. Second, the fact that farms tend to be larger in crop and livestock industries without subsidies suggests subsidies may have inhibited consolidation where they have been important. Third, truly tiny farms (half the total official number) often do not bother filing for subsidies, but this has no relationship with commercial farm size. Fourth, subsidies vary from year to year inversely with market price and yield, so subsidies jump in years with size declines as measured by sales or output.

Conceptual effects can be subtle. Subsidies that offset market variability may allow weaker farms to access capital not otherwise available. Payments based on historical production allow even farms with less productivity growth to remain viable longer, and subsidy income capitalized into farm land and other farm assets allow equity owners to survive when farms might otherwise consolidate.

The empirical evidence on the effects of farm programs and related policies on farm size distributions is mixed. A series of papers by Key and Roberts (for example, their 2007 article) used panel data from Census records and provided support for the hypothesis that farm subsidies encouraged farmland consolidation and made farms bigger. However, Census data did not allow researchers to identify effects of specific farm programs on different eligible crops and the various potential relationships with government payments, trade protection, and subsidized crop insurance. The major conclusions of Key and Roberts seem to be most heavily influenced by impacts on tiny farms that are of little interest for aggregate farm production or other issues of commercial agriculture. Key and Roberts also cannot account for the differences in subsidy and size across crop industries (MacDonald, Korb, and Hoppe 2013).

There is evidence that certain farm program features affect size in direct ways. Foltz (2004) showed that some very small New England dairy farms delayed exiting the industry in response to a targeted payment that raised the effective price of milk for farms in that region. Kirwan, Uchida, and White (2012) showed that elimination of the tobacco program that operated as a supply control cartel and restricted downstream marketing relationships allowed rapid farm consolidation that had been delayed by the operation of the program. Sneeringer and Key (2011) document how regulations that create size thresholds for more intensive and costly regulatory compliance have caused livestock operations to limit operational size to just below the threshold. Finally, the Food Safety Modernization Act of 2012 explicitly exempted small farms from some rules. One rationale for such exemptions is that large farms may have lower per unit costs of understanding and implementing regulatory requirements.

The answers as to how government farm subsidy programs affect farm size ultimately hinge on the degree of relative subsidy rates across competing commodities and farm sizes, the supply elasticities of the subsidized crops, and effects on such second-order factors as variability of returns, access to operating capital, and potential risk aversion.

Conclusion

Commercial agriculture in the United States is comprised of several hundred thousand farms, and these farms continue to become larger and fewer. The size of commercial farms is sometimes best-measured by sales, in other cases by acreage, and in still other cases by quantity produced of specific commodities, but for many commodities, size has doubled and doubled again in a generation. That does not mean that typical commercial farm operations are becoming large by any nonfarm corporate standard, or that there is any near-term prospect that these large firms will be able to exercise market power. For example, even as the typical herd size of dairy farms rises from 500 cows to 1,000 to 2,000, there will remain thousands of commercial farms operating the national milk cow herd of eight or nine million cows. The few dairy farms with 10,000 cows are located in several units in distinct locations and remain a small share of the relevant national and international market into which they deliver.

In a few farm commodity industries, multi-location operations and hierarchical management systems have emerged. In some commodity industries, such as eggs and some specialized fruits, tree nuts, and vegetables, farms have consolidated enough that most national production derives from fewer than 100 major producers who operate at several locations separated by hundreds or thousands of miles. These industries are characterized by close relationships between growing and marketing, seemingly smaller uncontrollable crop-to-crop variations, and therefore less costly monitoring of managerial outcomes.

In some industries, such as intensive animal feeding, farms are often operated as franchises in which farms are connected closely with larger processing and marketing firms through contractual relationships (MacDonald and McBride 2009). Many commodity industries have traditionally used contractual relationships between farms and processors or marketers to coordinate timing of shipments and commodity characteristics. For example, the processing tomato industry links growers and processors in annually negotiated contracts, and wineries work closely with contracted grape growers, often providing long-term guarantees to encourage vineyard development. Growth in farm size in these industries has occurred at roughly the same pace as for commodity industries with fewer contractual relationships. Economists do not yet have a good understanding of the relationships between contractual relationships, farm size patterns, and productivity, and this remains an area of active research.

Changes in farm size distributions and growth of farms seems closely related to technological innovations, managerial capability, and productivity. Opportunities for competitive returns from investing financial and human capital in farming hinge on applying managerial capability to an operation large enough to provide sufficient payoff. Farms with better managers grow, and these managers take better advantage of innovations in technology, which themselves require more technical and managerial sophistication. Farms now routinely use outside consultants for technological services such as animal health and nutrition, calibration and timing of fertilizers and pesticides, and accounting. The result is higher productivity, especially in reducing labor and land per unit of output. Under this scenario, agricultural research leads to technology that pays off most to more-capable managers who operate larger farms that have lower costs and higher productivity. The result is reinforcing productivity improvements.

Subsidy programs seem to be relatively unimportant in the evolution of farming in the United States. Farm sizes are growing, numbers of commercial farms are falling, and farm operations are transforming industries with and without commodity subsidies. In specific instances and for specific commodities, farm programs have affected the patterns of farm size and growth. But the broader swath of agricultural transformation is affected, not by subsidies, which in recent years amount to about 5 percent of US farm revenue, but by markets that reward innovation.

Agriculture continues to evolve rapidly in the United States. Recent research has been stimulated by annual US Department of Agriculture surveys of relatively large, repeated cross sections of farms that can be appropriately weighted to be representative of particular commodity industries. These data along with repeated individual farm records from the US Census of Agriculture are providing a rich source of information that will allow us to better understand how and why farm size distributions are changing and what this means for the agricultural landscape, public policy, and agricultural productivity.

■ *The author thanks research assistants John Bovay, Eric Coughlin, Georgi Gabrielyan, and Jisang Yu. Rick Barichello provided useful comments. I especially appreciate correspondence and conversations with James MacDonald of the Economic Research Service, US Department of Agriculture. The work of MacDonald and his colleagues has brought substantial new evidence and analysis to bear on the issues considered in this article.*

References

Allen, Douglas W., and Dean Lueck. 1998. "The Nature of the Farm." *Journal of Law and Economics* 41(2): 343–86.

Alston, Julian M., Matthew A. Andersen, Jennifer S. James, and Philip G. Pardey. 2010. *Persistence Pays: U.S. Agricultural Productivity Growth and the Benefits from Public R&D Spending.* New York: Springer.

Alvarez, Antonio M., and Carlos Arias. 2003. "Diseconomies of Size with Fixed Managerial

Ability." *American Journal of Agricultural Economics* 85(1): 134–42.

Angelini, Paolo, and Andrea Generale. 2008. "On the Evolution of Firm Size Distributions." *American Economic Review* 98(1): 426–38.

Brown, Jason P., and Jeremy G. Weber. 2013. *The Off-Farm Occupations of U.S. Farm Operators and Their Spouses.* Economic Information Bulletin 117, Economic Research Service, US Department of Agriculture. September.

Cabral, Luís, and José Mata. 2003. "On the Evolution of the Firm Size Distribution: Facts and Theory." *American Economic Review* 93(4): 1075–90.

Dimitri, Carolyn, Anne Effland, and Neilson Conklin. 2005. *The Twentieth Century Transformation of U.S. Agriculture and Farm Policy.* Economic Information Bulletin No. 3, Economic Research Service, US Department of Agriculture. June.

Eastwood, Robert Michael Lipton, Andrew Newell. 2010. "Farm Size." Chap. 65 in *Handbook of Agricultural Economics,* vol. 4, edited by Prabhu L. Pingali and Robert E. Evenson, 3323–97. Elsevier.

Economic Research Service. United States Department of Agriculture (USDA). February 2013. *USDA Long-Term Agricultural Projection Tables.* http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1192.

Economic Research Service, United States Department of Agriculture (USDA). 2013. "ARMS Farm Financial and Crop Production Practices/Tailored Reports." Results from the Agricultural Resource Management Survey (ARMS). http://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices/tailored-reports.aspx#P0288e93c65b24def9c83f9336087cd84_5_66iT0R0T0R0x5.

Evans, David S. 1987. "Tests of Alternative Theories of Firm Growth." *Journal of Political Economy* 95(4): 657–74.

Evans, David S., and Boyan Jovanovic. 1989. "An Estimated Model of Entrepreneurial Choice under Liquidity Constraints." *Journal of Political Economy* 97(4): 808–27.

Evans, David S., and Linda S. Leighton. 1989. "Empirical Aspects of Entrepreneurship." *American Economic Review* 79(3): 519–35.

Feder, Gershon. 1985. "The Relation between Farm Size and Farm Productivity: The Role of Family Labor, Supervision and Credit Constraints." *Journal of Development Economics* 18(2–3): 297–313.

Foltz, Jeremy D. 2004. "Entry, Exit, and Farm Size: Assessing an Experiment in Dairy Price Policy." *American Journal of Agricultural Economics* 86(3): 594–604.

Fuglie, Keith O., James M. MacDonald, and Eldon Ball. 2007. "Productivity Growth in U.S. Agriculture." Economic Brief No. EB-9, United States Department of Agriculture, Economic Research Service, September.

Gardner, Bruce L. 1992. "Changing Economic Perspectives on the Farm Problem." *Journal of Economic Literature* 30(1): 62–101.

Gardner, Bruce L. 2002. *American Agriculture in the Twentieth Century: How It Flourished and What It Cost.* Harvard University Press.

Glauber, Joseph W. 2013. "The Growth of the Federal Crop Insurance Program, 1990–2011." *American Journal of Agricultural Economics* 95(2): 482–88.

Hall, Bronwyn H. 1987. "Empirical Analysis of the Size Distribution of Farms: Discussion." *American Journal of Agricultural Economics* 69(2): 486–87.

Hendricks, Nathan P., and Daniel A. Sumner. Forthcoming. "The Effects of Policy Expectations in Crop Supply, with an Application to Base Updating." *American Journal of Agricultural Economics.*

Jovanovic, Boyan. 1982. "Selection and the Evolution of Industry." *Econometrica* 50(3): 649–70.

Key, Nigel D., and Michael J. Roberts. 2007. "Commodity Payments, Farm Business Survival, and Farm Size Growth." Economic Research Report No. EER-51, United States Department of Agriculture, November.

Kirwan, Barrett E., Shinsuke Uchida, and T. Kirk White. 2012. "Aggregate and Farm-Level Productivity Growth in Tobacco: Before and After the Quota Buyout." *American Journal of Agricultural Economics* 94(4): 838–53.

Kislev, Yoav, and Willis Peterson. 1982. "Prices, Technology and Farm Size." *Journal of Political Economy* 90(3): 578–95.

Lucas, Robert E., Jr. 1978. "On the Size Distribution of Business Firms." *Bell Journal of Economics* 9(2): 508–23.

Luttmer, Erzo G. J. 2011. "On the Mechanics of Firm Growth." *Review of Economic Studies* 78(3): 1042–68.

MacDonald, James M., and William D. McBride. 2009. *The Transformation of U.S. Livestock Agriculture: Scale, Efficiency and Risks.* Economic Information Bulletin No EIB-43. US Department of Agriculture, Economic Research Service, January.

MacDonald, James M., Penni Korb, and Robert A. Hoppe. 2013. *Farm Size and the Organization of U.S. Crop Farming.* Economic Research Report No. ERR-52, Economic Research Service, US Department of Agriculture. August.

McBride, William D., and Nigel Key. 2013. "U.S. Hog Production from 1992 to 2009: Technology, Restructuring, and Productivity Growth." Economic Research Report No. ERR-158, US Department of Agriculture, Economic Research Service, October.

Melhim, Almuhand, Erik J. O'Donoghue, and C. Richard Shumway. 2009. "Do the Largest Firms Grow and Diversify the Fastest? The Case of U.S. Dairies." *Review of Agricultural Economics* 31(2): 284–302.

Mosheim, Roberto, and C. A. Knox Lovell. 2009. "Scale Economies and Inefficiency of U.S. Dairy Farms." *American Journal of Agricultural Economics* 91(3): 777–94.

National Agricultural Statistics Service, United States Department of Agriculture. 1994. *California Agricultural Statistics 1993,* November, pg. 16. http://www.nass.usda.gov/Statistics_by_State/California/Publications/California_Ag_Statistics/Reports/1993cas-all.pdf.

National Agricultural Statistics Service, United States Department of Agriculture. 1994–2011. "Agricultural Statistics Annual." United States Department of Agriculture. http://www.nass.usda.gov/Publications/Ag_Statistics/index.asp (accessed March 7, 2013).

National Agricultural Statistics Service, United States Department of Agriculture, 2012. *California Agricultural Statistics, 2011 Crop Year.* October 31. http://www.nass.usda.gov/Statistics_by_State/California/Publications/California_Ag_Statistics/Reports/2011cas-all.pdf.

O'Donoghue, Erik J., Robert A. Hoppe, David E. Banker, Robert Ebel, Keith Fuglie, Penni Korb, Michael Livingston, Cynthia Nickerson, and Carmen Sandretto. 2011. *The Changing Organization of U.S. Farming.* Economic Information Bulletin No. EBI-88, Economic Research Service, US Department of Agriculture. December.

O'Donoghue, Erik J., Robert A. Hoppe, David Banker, and Penni Korb. 2009. *Exploring Alternative Farm Definitions: Implications for Statistics and for Program Eligibility.* Economic Information Bulletin No. EBI-49, Economic Research Service, US Department of Agriculture. March.

Risk Management Agency, United States Department of Agriculture. 2013. "Program Costs and Outlays." http://www.rma.usda.gov/aboutrma/budget/costsoutlays.html.

Roberts, Michael J., and Nigel Key. 2008. "Agricultural Payments and Land Concentration: A Semiparametric Spatial Regression Analysis." *American Journal of Agricultural Economics* 90(3): 627–43.

Sneeringer, Stacy Ellen, and Nigel Key. 2011. "Effects of Size-Based Environmental Regulations: Evidence of Regulatory Avoidance." *American Journal of Agricultural Economics* 93(4): 1189–1211.

Sumner, Daniel A. 1985. "Farm Programs and Structural Issues." In *U.S Agricultural Policy: The 1985 Farm Legislation*, edited by Bruce L. Gardner, AEI Symposia series no. 85C, pp. 283–320. Washington, DC.

Sumner, Daniel A., and James D. Leiby. 1987. "An Econometric Analysis of the Effects of Human Capital on Size and Growth among Dairy Farms." *American Journal of Agricultural Economics* 69(2): 465–70.

Sumner, Daniel A., Julian M. Alston, and Joseph W. Glauber. 2010. "Evolution of the Economics of Agricultural Policy." *American Journal of Agricultural Economics* 92(2) 403–23.

Sumner, Daniel A., and Christopher A. Wolf. 2002. "Diversification, Vertical Integration, and the Regional Pattern of Dairy Farm Size." *Review of Agricultural Economics* 24(2): 442–57.

Tolley, George S. 1970. "Management Entry into U.S. Agriculture." *American Journal of Agricultural Economics* 52(4): 485–93.

US Department of Agriculture. 1992, 1997, 2002, 2007. "Census of Agriculture." http://www.agcensus.usda.gov/Publications/Historical_Publications/index.php (accessed March 7, 2013).

US Small Business Administration. Undated. "What is SBA's Definition of a Small Business Concern?" United States Small Business Administration. http://www.sba.gov/content/what-sbas-definition-small-business-concern (accessed March 9, 2013).

White, T. Kirk, and Robert A. Hoppe. 2012. *Changing Farm Structure and the Distribution of Farm Payments and Crop Insurance.* Economic Information Bulletin No. EIB-91, Economic Research Service, US Department of Agriculture. February.

Wolf, Christopher A., and Daniel A. Sumner. 2001. "Are Farm Size Distributions Bimodal? Evidence from Kernel Density Estimates of Dairy Farm Size Distributions." *American Journal of Agricultural Economics* 83(1): 77–88.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 167–188

From Sick Man of Europe to Economic Superstar: Germany's Resurgent Economy†

Christian Dustmann, Bernd Fitzenberger, Uta Schönberg, and Alexandra Spitz-Oener

In the late 1990s and into the early 2000s, Germany was often called "the sick man of Europe" (for example, *Economist* 2004), a phrase usually attributed to comments by Czar Nicholas I of Russia about the troubles faced by the Ottoman Empire in the mid-19th century. Indeed, Germany's economic growth averaged only about 1.2 percent per year from 1998 to 2005, including a recession in 2003, and unemployment rates rose from 9.2 percent in 1998 to 11.1 percent in 2005 (according to World Bank data). Today, after the Great Recession, Germany is described as an "economic superstar" (for example, in the movie "Made in Germany: Europe's Economic Superstar," http://films.com/ItemDetails.aspx?TitleId=29218). Germany's number of total unemployed fell from 5 million in 2005 to about 3 million in 2008, and its unemployment rate had declined to 7.7 percent in 2010 (according to data from Germany's Federal Employment Agency, the Bundesagentur für Arbeit). In contrast to most of its European neighbors and the United States, Germany experienced almost no increase in unemployment during the Great Recession, despite a sharp decline in GDP in 2008 and 2009 (an episode discussed

■ *Christian Dustmann is Professor of Economics, University College London, London, United Kingdom. Bernd Fitzenberger is Professor of Statistics and Econometrics, University of Freiburg, Freiburg, Germany. Uta Schönberg is Associate Professor of Economics, University College London, London, United Kingdom. Alexandra Spitz-Oener is Professor of Economics, Humboldt University Berlin, Berlin, Germany. Fitzenberger is also Research Associate at ZEW Mannheim, and Schönberg and Spitz-Oener are also Research Associates at the IAB, Nuremberg, Germany. Their email addresses are c.dustmann@ucl.ac.uk, bernd.fitzenberger@vwl.uni-freiburg.de, u.schoenberg@ucl.ac.uk, and alexandra.spitz-oener@wiwi.hu-berlin.de.*

† To access the Appendix, visit
http://dx.doi.org/10.1257/jep.28.1.167

doi=10.1257/jep.28.1.167

in Möller 2010; Burda and Hunt 2011). Germany's exports reached an all-time record of $1.738 trillion in 2011, which is roughly equal to half of Germany's GDP, or 7.7 percent of world exports. Even the euro crisis seems not to have been able to stop Germany's strengthening economy and employment.

How did Germany, with the fourth-largest GDP in the world (after the United States, China, and Japan) transform itself from "the sick man of Europe" to an "economic superstar" in less than a decade? One common answer points to a series of legislative labor market reforms that started in the mid 2000s, the so-called "Hartz reforms." Another explanation focuses on the evolution of Germany's economy and trade balance in the context of the eurozone. However, we will argue that these factors did not play a decisive role for the transformation of the German economy, namely the restructuring of its labor market and the increase in competitiveness that has helped German exports. We instead present evidence that the specific governance structure of the German labor market institutions allowed them to react flexibly in a time of extraordinary economic circumstances, and that this distinctive characteristic of its labor market institutions has been the main reason for Germany's economic success over the last decade.[1]

We begin by arguing that the evolution of Germany's per unit labor costs—that is, labor costs relative to productivity—in both the manufacturing sector and the other sectors in the economy has played an important role in the favorable evolution of German tradable manufacturing industry. We then investigate the mechanisms that allowed for wage restraints and the dramatic decrease in real wages at the lower end of the wage distribution.

The specific feature of the German system of industrial relations that we stress is that it is not rooted in legislation, but instead is laid out in contracts and mutual agreements between the three main actors in Germany: employer associations, trade unions, and works councils. The institutional setup of this system, which is dominated by industry-wide wage bargaining, remained basically unchanged. However, many indicators demonstrate that it did change in the way it operates. For example, the share of German workers covered by any kind of union agreement has sharply declined, and the number of firm-level deviations from industry-wide union agreements has sharply increased since the mid 1990s. Overall, these gradual changes within the system led to an unprecedented decentralization of the wage-setting process from the industry level to the firm level. Alternatively, one may refer to this process as an increasing localization of Germany's industrial relations.

The decentralization in wage setting in Germany is in contrast to many of its neighbors where the statutory minimum wage is often high (relative to productivity), where union wages and work hour regulations apply to all firms in the

[1] Our argument is similar in spirit to that of Carlin and Soskice (2008, 2009), who argue that it is restructuring by Germany's private sector, using traditional German institutions based on employer-worker cooperation, and not government labor market and welfare state reforms that are to be credited for the German recovery.

industry, and where institutional change therefore requires broad consensus along the political spectrum.

We then turn to a discussion of why Germany's labor market experience has been so distinctive within continental Europe. On the one side, the fall of the Berlin Wall in 1989 and the dramatic cost of reunification burdened the German economy in an unprecedented way, leading to a prolonged period of dismal macroeconomic performance. On the other side, it gave German employers access to neighboring East European countries that were formerly locked away behind the Iron Curtain, and that were characterized by low labor cost, yet stable institutions and political structures. These factors changed the power equilibrium between employer and employee associations and forced the latter to respond in a far more flexible way than many would ever have expected. Finally, we discuss the relationship between our analysis of the flexibility of Germany's labor market institutions and two other events: Germany's Hartz reforms of 2003 and the arrival of the euro in 1999.

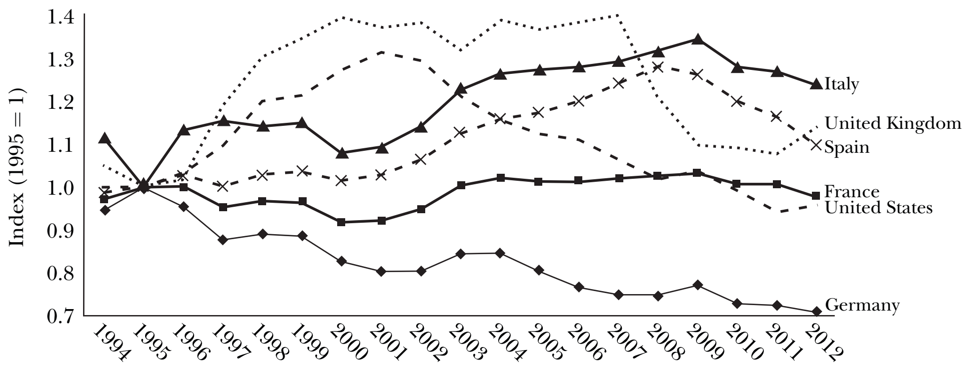
How Did Germany Improve its Competitiveness?

Relative Unit Labor Cost

In Figure 1, we plot the "relative unit labor costs" for a country's overall economy adjusted for the changing composition of the markets in which it competes, for a selection of countries, in dollar terms. This index is computed by the OECD based on year-to-year changes of unit labor costs and shows the relative change in the unit labor costs over time (normalized to 1995) translated into US dollars at the current exchange rate compared to a weighted average of a country's trading partners. The weights of the trading partners adjust annually to changes in trading patterns. An increase in this index indicates a deterioration of the competitive position. A drop in this index—that is, an improvement in competitiveness—is caused by some combination of three factors: 1) a decrease in the wage per worker (or per hour); 2) an increase in productivity (per worker or per hour); and 3) a nominal depreciation of a country's foreign exchange rate.

Since 1995, Germany's competitive position has persistently improved, while the competitiveness of some of its main European trading partners has deteriorated (Spain and Italy) or remained close to the 1995 position (France). The competitiveness of the United Kingdom has likewise deteriorated, although it improved dramatically between 2007 and 2009 due to the sharp depreciation of the British pound against other currencies. The US economy also lost competitiveness relative to Germany in the late 1990s as the US dollar appreciated in value relative to European currencies, but improved consistently after the 2001 recession, partly achieved through a dollar depreciation (for instance, while the euro/dollar exchange rate was around 1 in 2001, it had depreciated to 0.8 in 2009). However, Germany's gains in competitiveness with regard to France, Italy, and Spain cannot be due to currency depreciation (and in fact the euro appreciated relative to the currency of most trading partners), because these countries all share the euro, and so it must

Figure 1
Evolution of Competition-Weighted Relative Unit Labor Costs, Selected Countries, 1994–2012



Source: OECD Economic Indicators.
Notes: This index accumulates the annual change in the relative unit labor costs of country i compared to a weighted average of its main trading partners where labor costs are translated into dollars and the weights are adjusted annually to the change in trade pattern. The annual change in logs is calculated as $\Delta\log(RULC_{it}) = \Delta\log(ULC_{it}\, e_{it}) - \sum_{j\neq i} g_{ij}^{t-1}\, \Delta\log(ULC_{jt}\, e_{jt})$ where $ULC_{it} = (w_{it}L_{it})/Y_{it}$ is the unit labor cost for country i in period t, computed as the total wage bill $w_{it}L_{it}$ divided by the value added of the country's industry Y_{it}. The unit labor costs are translated into US dollars using the exchange rate e_{it}. Both the unit labor costs and the exchange rates are defined as index relative to some base year. The weighting scheme g_{ij}^{t-1} takes account of the structure of competition in both export and import markets of the goods sector of those countries, and it adjusts on a year-by-year basis. See OECD *Economic Outlook* (2012, Issue 2, No. 92) and OECD Economic Outlook Sources and Methods (http://www.oecd.org/eco/sources-and-methods) for details on the method of calculation.

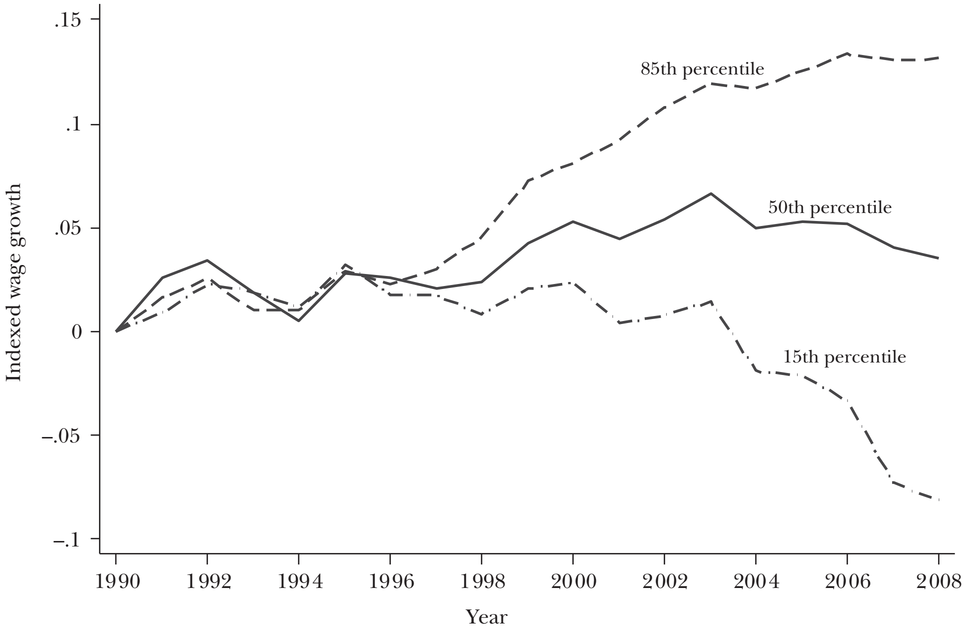
have arisen because German wages grew at a slower pace than productivity relative to these other eurozone countries.

Wage Trends and Wage Inequality

Figure 2 shows the evolution of real wages in West Germany since 1990. The figure illustrates the dramatic development in wage inequality in West Germany over the past 15 years or so (Dustmann, Ludsteck, and Schönberg 2009; see also Antonczyk, Fitzenberger, and Sommerfeld 2010; Card, Heining, and Kline 2013).[2] Real wages at the 15th percentile fell dramatically from the mid 1990s onwards. From the early 2000s onwards, median real wages started to fall, and only wages at the top of the distribution continued to rise. Notice that all wage figures that we report stand for West Germany (although, henceforth, we refer to them as "Germany"), because developments in East Germany are strongly affected by the transition after German unification.

[2] Details on the wage data are in Appendix A available online with this journal at http://e-jep.org.

Figure 2
Indexed Wage Growth of the 15th, 50th, 85th Percentiles, West Germany, 1990–2008



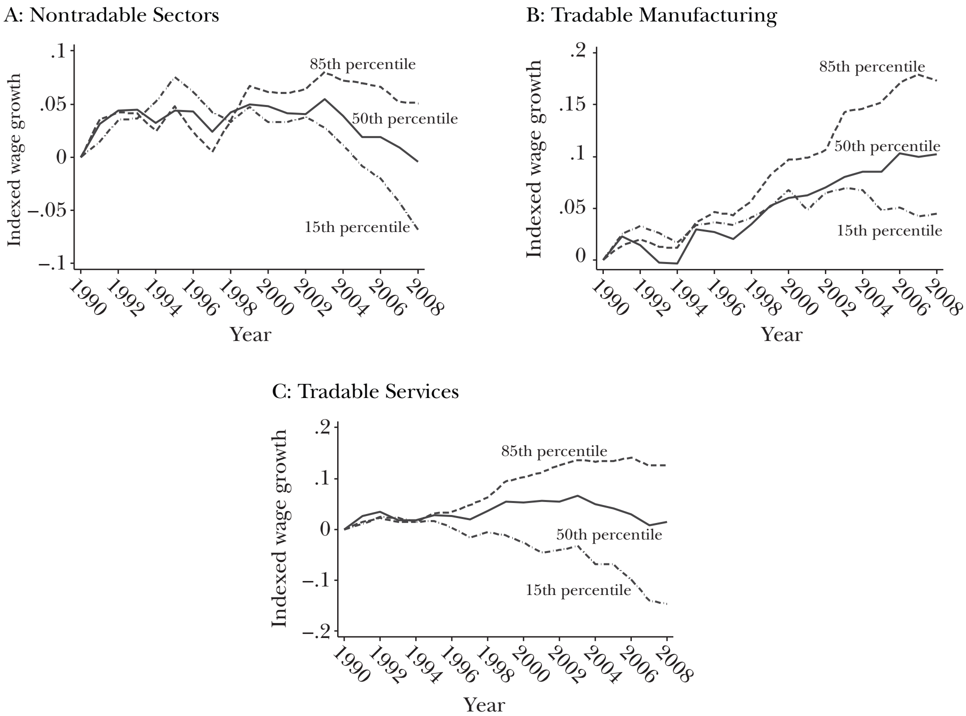
Notes: Calculations based on SIAB Sample for West German Full-Time Workers between 20 and 60 years of age. The figure shows the indexed (log) real wage growth of the 15th, 50th, and 85th percentiles of the wage distribution, with 1990 as the base year. Nominal wages are deflated using the consumer price index (1995 = 100) provided by the German Federal Statistical Office.

If the increase in wage inequality and the modest growth in wages overall—and in particular the dramatic decline in real wages at the bottom of the wage distribution—has contributed to the favorable evolution of unit labor costs in Germany relative to the United States and other eurozone countries, then one should expect this development to have been particularly pronounced in the tradable manufacturing sector—the backbone of the German exporting industries accounting for 80 percent of German exports. This insight turns out to hold true, but in an unexpected way.

To further explore the increase in wage inequality, we classify sectors with export volumes below the 25th percentile of the distribution of export volumes in 1995 as "nontradable sectors," and those with export volumes above this threshold as "tradable sectors." "Tradable manufacturing" are all those tradable sectors that belong to the manufacturing sector, and "tradable services" are all other tradable sectors.[3] Figure 3 breaks down the evolution of real wages along the wage

[3] Details on the construction of these categories can be found in Appendix A, available with this paper at http://e-jep.org.

Figure 3
Indexed Wage Growth of the 15th, 50th, 85th Percentiles, West Germany, by Sectors, 1990–2008



Notes: Calculations based on SIAB Sample for West German Full-Time Workers between 20 and 60 years of age. The figures show the indexed (log) real wage growth of the 15th, 50th, and 85th percentiles of the wage distribution, with 1990 as the base year. Nominal wages are deflated using the consumer price index (1995 = 100) provided by the German Federal Statistical Office. Panel A shows the evolution of these figures for the nontradable sectors, panel B for tradable manufacturing, and panel C for tradable services. We classify sectors with export volumes below the 25th percentile of the distribution of export volumes in 1995 as "nontradable sectors", and those with export volumes above this threshold and that belong to the manufacturing sector as "tradable manufacturing." The sectors above this threshold that do not belong to the manufacturing sector are classified as "tradable services."

distribution separately for the nontradable sector, tradable manufacturing sector (henceforth denoted as manufacturing), and tradable services sector. By this measure, real wages in the manufacturing sector rose at all percentiles of the wage distribution until the mid 2000s and afterwards continued to rise at the median and the 85th percentile. Germany's real wages in the nontradable sector hardly increased at all at any part of the wage distribution during the 1990s and started to decline from the early 2000s onwards even at the 85th percentile, but particularly so at the 15th percentile. The sharpest increase in inequality occurred in the tradable service sector, where between 1990 and 2008 real wages did not show an increase at the median, increased by 12 percent at the 85th percentile and declined by almost

15 percent at the 15th percentile. At first glance, these figures do not seem to lend support to the hypothesis that wage restraint in the manufacturing sector was an important factor in improving competitiveness in that sector.

Exports, Tradable Manufacturing, and Domestic Inputs

The end product in manufacturing, however, contains a large share of inputs produced in other sectors: in Germany, the value added in manufacturing is only roughly one-third of the value of the end product, with the remainder of value added being contributed through inputs from other industries, either domestically or from abroad (the literature so far has focused on Germany's imports of intermediate products from abroad, see Geishecker 2006; Sinn 2006; OECD 2007, chap. 3; OECD 2012, chap. 3). Hence, the manufacturing sector may have benefited from low wages in other domestic sectors and from cheap imports from abroad. In addition, Germany's manufacturing sector may have experienced increases in productivity which exceeded the increases in wages in the manufacturing sector.

More detailed evidence suggests that both of these factors may be at play.[4] In Germany, the manufacturing sector comprised 21.6 percent of all jobs in 1995, but 17.7 percent of all jobs in 2007, while the value added of this sector (in current prices) remained essentially unchanged at 22.8 percent of all value added in 1995 compared with 22.7 percent of value added in 2007. This pattern suggests larger productivity increases in the manufacturing sector than in the other sectors, where employment shares increased over the same period, with value added remaining roughly constant. This pattern is not uncommon across high-income countries.[5] However, the share of manufacturing in output value (value of final products), as opposed to value added, rose steadily from 35 percent of all output in 1995 to 39.3 percent of all output in 2007. This pattern reflects that the manufacturing sector indeed relies to an increasing extent on inputs from other domestic sectors and on imported inputs (because the share in final products has increased while the share in value added has remained the same), and may thus have benefited from the low wage growth in other domestic sectors and from cheaper imports.

Digging down into the more detailed data, shown in Table 1, the value of inputs over the value of output is nearly twice as high in manufacturing as in the other two sectors (66.1 percent in 1995 versus 37.8 percent in the tradable service sector) and this share increased by about 7 percentage points to 72.9 percent in 2007. The share of domestic inputs remained constant over the same period at about 51 percent. Thus, the increase in the share of inputs used by Germany's manufacturing sector, relative to the output value in that sector, is driven by increased

[4] See Table 1 and Table A1 in Appendix C available online with this paper at http://e-jep.org for details and data sources.

[5] Pilat, Cimper, Olsen, and Webb (2006) point out that the relatively fast productivity growth in manufacturing is associated with relative declines of the prices for manufacturing products (this is Baumol's cost disease). Thus, shares in value added at current prices understate the share of value added at constant prices in manufacturing to total value added at constant prices, which makes it remarkable that manufacturing in Germany has retained its share in value added at current prices.

Table 1
Evolution of the Share of Value of Total Inputs and Domestic Inputs over the Value of Output, Overall and by Sector, 1995–2007

	Overall	*Nontradable sectors*	*Tradable manufacturing*	*Tradable services*
Panel A: Value of Total Inputs/Output Value				
1995	48.2%	39.9%	66.1%	37.9%
2000	51.0%	37.9%	70.1%	41.4%
2007	53.2%	38.2%	72.9%	41.6%
Panel B: Value of Domestic Inputs/Output Value				
1995	39.8%	35.3%	51.7%	32.4%
2000	40.3%	32.2%	51.7%	34.8%
2007	40.5%	32.1%	51.2%	34.2%
Panel C: Value of Domestic Inputs/Value of Total Inputs				
1995	82.6%	88.3%	78.1%	85.6%
2000	79.0%	84.9%	73.7%	84.0%
2007	76.1%	83.9%	70.3%	82.2%

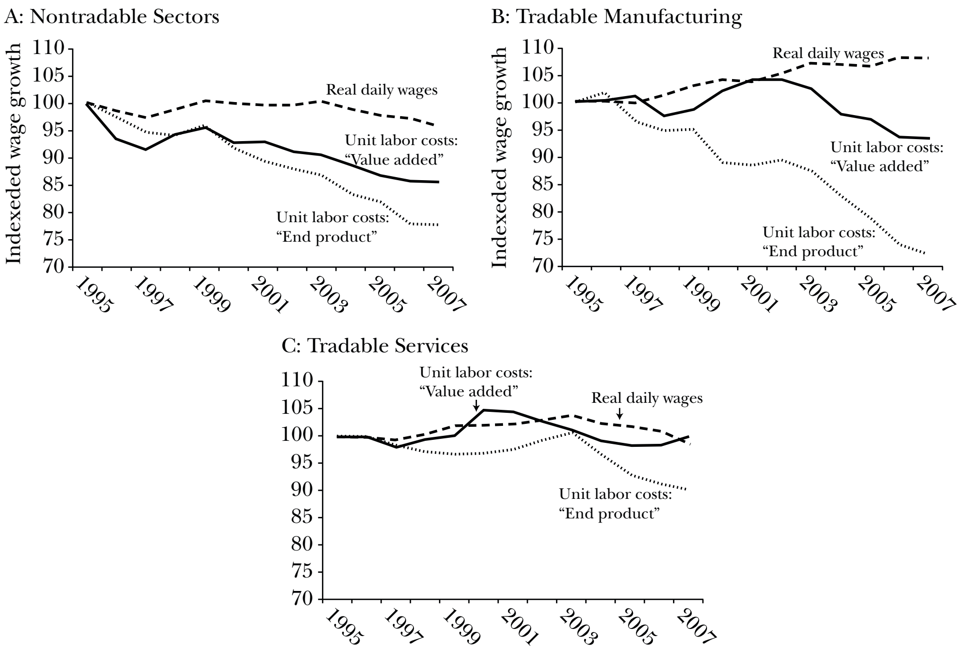
Notes: Calculations based on input-output statistics from the German Statistical Office (Fachserie 18, Reihe 2, Years: 1995–2007). We classify sectors with export volumes below the 25th percentile of the distribution of export volumes in 1995 as "nontradable sectors" and those with export volumes above this threshold and that belong to the manufacturing sector as "tradable manufacturing." The sectors above this threshold that do not belong to the manufacturing sector are classified as "tradable services."

use of inputs from abroad relative to inputs from domestic industries. However, even in 2007, 70 percent of overall inputs in Germany's manufacturing sector were domestically produced. Thus, the argument that Germany's manufacturing sector has become nothing more than an assembly place for foreign produced inputs (for example, Sinn 2006) is unjustified. In fact, while German manufacturing has made increasing use of imported inputs, the share of domestic inputs in manufacturing final output value had remained high and relatively stable between 1995 and 2007.

To what extent have Germany's domestic inputs contributed to competitiveness in its export-oriented manufacturing sector and the two other sectors? In Figure 4, we plot the evolution of unit labor costs in the three sectors, where industries are weighted with respect to their exports for the two tradable sectors.[6] When computing unit labor costs, we first consider only the value added in the sector, as denoted by solid lines in Figure 4. We then consider final output value in the sector, which is the sum of value added in the sector and all inputs into the sector denoted by dotted lines in Figure 4. This index (Unit Labor Costs: "End Products") incorporates gains in competitiveness in a sector due to the usage of inputs from other domestic sectors. We also plot median real wages, adjusted using Germany's Consumer Price Index, for the three sectors. While real wage growth in the manufacturing sector is relatively modest,

[6] Details on how unit labor costs are calculated can be found in Appendix A and in Appendix C, which are available online with this article at http://e-jep.org.

Figure 4
Evolution of Real Daily Wages and Unit Labor Costs by Sector, 1995–2007



Notes: The figures show indexed real mean daily wages by sector (base year 1995 = 100). Nominal wages are deflated using the consumer price index (1995 = 100) provided by the German Federal Statistical Office. The data underlying these indices are in columns (1), (4), and (7) of Table A2. The figure also shows indexed unit labor costs both "Value added" and "End product" by sector. The data underlying these indices are in columns (2), (3), (5), (6), (8), and (9) of Table A2 of the online Appendix. Table A2 includes a detailed description of data and methods of calculations.

at about 8.2 percent over the 11-year period, in the other two sectors average wages fell in real terms by 1.2 and 4.1 percent, respectively, over this time period.

As visible in the figure, domestic unit labor cost for total production in manufacturing, taking account of inputs produced in other sectors ("end products"), declined far more rapidly than unit labor costs in value added—a decline that cannot be explained by the increase in the share of imported inputs in total output value. Moreover, unit labor costs in end products start to decline at the start of the observation period in 1995, while unit labor costs in value added decrease rapidly only from 2003 onwards when mean wages, and in particular wages at the 15th percentile of the wage distribution start to decrease in real terms (as shown earlier in Figure 3).

Thus, Germany's manufacturing sector improved competitiveness in several ways. First, manufacturing drew on inputs from domestically provided nontradable and especially tradable services, where real wages fell between 1995 and 2007. Second, the decline in unit labor costs, coupled with the increase in mean real wages in manufacturing, implies that productivity increases in the manufacturing sector have outpaced wage increases in that sector. In comparison, total unit labor costs fell less in the nontradable sectors (minus 22.2 percent) and much less in the tradable services (minus 9.7 percent), even though nominal wages grew much less in these two sectors compared to tradable manufacturing. Note also that productivity increases in the manufacturing sector have exceeded the increases in the two other sectors. Finally, to increase the competitiveness of its own final products, the manufacturing sector has made increased use of trade integration with Eastern European countries through inputs imported from abroad, and far more so than other European countries. These inputs made up 14.5 percent of total output in the manufacturing sector in 1995 and 21.5 percent in 2007. Calculating the outsourcing indicator suggested by Egger and Egger (2003, p. 642) for Germany, France, and Italy regarding imported inputs from Poland, Hungary, and the Czech and the Slovak Republics, using data from the OECD Input-Output-Tables (at http://www.oecd.org/trade/input-outputtables.htm) and OECD International Trade and Balance of Payments Statistics (at http://www.oecd.org/std/its/), shows that in the year 2000, imported inputs from these four countries amounted to about 8.5 percent of inputs in Germany, compared to 2.5 percent in Italy and 1.9 percent in France (relative to GDP).

The Increase in Competitiveness and Germany's Labor Market Institutions

The movements in German wages, within and across sectors, belie the common belief that Germany's labor market institutions are overly rigid. Instead, we argue that the specific governance structure of the German system of industrial relations offers various margins of flexibility. In the early to mid 1990s, these institutions allowed for an unprecedented increase in the decentralization (localization) of the process that sets wages, hours, and other aspects of working conditions, from the industry- and region-wide level to the level of the single firm or even the single worker, which in particular helped to bring down wages at the lower end of the wage distribution. This decentralization took place even though the institutional setup of the dominating system of industry-wide wage bargaining basically remained unchanged.

The specific feature which we stress here is that the governance structure of the German system of industrial relations is not rooted in legislation and is not governed by the political process, but instead is laid out in contracts and mutual agreements between the three main labor market parties: trade unions, employer associations, and works councils (the worker representatives who are typically present in medium-sized and large firms).[7] For this reason, Germany was in the position to react in an unprecedented way to the challenges of the early 1990s.

[7] Works councils have to be set up in establishments with more than five employees when demanded so by the employees. About 92 percent of employees in establishments that have more than 50 employees work in establishments with a works council, but only 18 percent of employees in establishments that are smaller (Addison. Schnabel, and Wagner 1997; Beckmann, Föhr, and Kräkel 2010).

The principle of autonomy of wage bargaining is laid down in the German constitution and implies that negotiations take place without the government directly exerting influence. As such, Germany has had no statutory minimum wage imposed by the political process over the period we study. Rather, an elaborate system of wage floors is negotiated periodically between trade unions and employer associations, typically at the industry and regional level.

This model of industrial relations has been very successful in Germany, where negotiation with unions and participation of work councils in decision-making processes are widely regarded as an important cornerstone in furthering common interests and even improving productivity. As a consequence, negotiations are usually far more consensus-based and less confrontational than in other countries. For example, Germany lost on average 11 days of work each year per 1,000 employees by strikes and lock-outs between 1991 and 1999, but only five days per 1,000 employees between 2000 and 2007. These figures for the earlier and later time period compare to 40 and 32 days per 1,000 employees in the United States, 30 and 30 days in the United Kingdom, 73 and 103 days in France, 158 and 93 days in Italy, and 220 and 164 days in Canada (Lesch 2009).

Germany's culture of common interest is dissimilar to the view about worker representations commonly held in the United States. A recent US example is the attempt of the management of the German company Volkswagen to introduce a works council at its Chattanooga plant in Tennessee. While the participation of works councils in management decisions is considered by Volkswagen as a cornerstone of successful firm policy that helps furthering common interests, Tennessee Governor Bill Haslam has been outspoken in opposing any union formation at the plant, fearing that it endangers the state's effort to attract investment (Greenhouse 2013). A key difference between US and German labor market institutions lies in the fact that a works council in Germany elected by the employees does not have to be a union representative (although in practice the majority of works councils are union representatives), while the installation of a works council in a US firm automatically involves the firm becoming unionized. Thus, works councils in Germany may act in greater independence from a union if the survival of their firm is at stake.

Unions and Employer Associations

In Germany, contractual agreements between unions and employer associations are negotiated either on the region-industry level or on the firm level. In addition to wages, working time regulations are an important component of the negotiations.

A distinguishing feature from US labor market institutions is that the recognition of trade unions in Germany is at the discretion of the firm, and union contracts cover only the workers in firms that recognize the relevant sectoral wage bargaining (union) contract—regardless of whether the worker is a union member (for discussion, see OECD 2004; Dustmann and Schoenberg 2009; Fitzenberger, Kohn, and Lembcke 2013). Also, German firms that once recognized the union contracts can later opt out at their own discretion. Even within union wage contracts negotiated at the industry level, there is scope for wage flexibility at the firm level through so-called "opening" or "hardship" clauses, provided that workers' representatives agree (for example, Hassel 1999; Hassel and Rehder 2001; Carlin and Soskice 2009; Brändle, Heinbach, and Meier 2011; Bispinck, Dribbusch, and Schulten 2010). After opting out of a collective agreement, firms still have to pay wages for the incumbent employees according to the collective agreement until a new agreement at the firm level has been reached, but they do not have to honor new negotiated wage increases and the firm need not follow the old collective agreements for new hires. Thus, over time a firm may be able to lower wage costs considerably by opting out of the union contract—provided its employees accepted this.

After 1995, there was indeed a dramatic decline in union coverage in Germany. This decline is almost entirely driven by a decline in industry-wide agreements.[8] From 1995 to 2008, the share of employees covered by industry-wide agreements fell from 75 to 56 percent, while the share covered by firm-level agreements fell from 10.5 to 9 percent. The percentage of German workers that were not covered by an agreement in 1995–1997 was highest in the tradable services (22 percent), as compared to tradable manufacturing (9.8 percent) and nontradables (12 percent). By 2006–2007, noncoverage had sharply increased in all three sectors to 40, 27, and 32 percent in the tradable services, manufacturing, and nontradables respectively, and this share continued to rise. By 2010, according to the German Structure of Earnings Survey, 41 percent of all employees in firms with at least 10 employees in the sectors Manufacturing, Mining, and Services are not covered by any collective wage agreement (StaBu 2013).

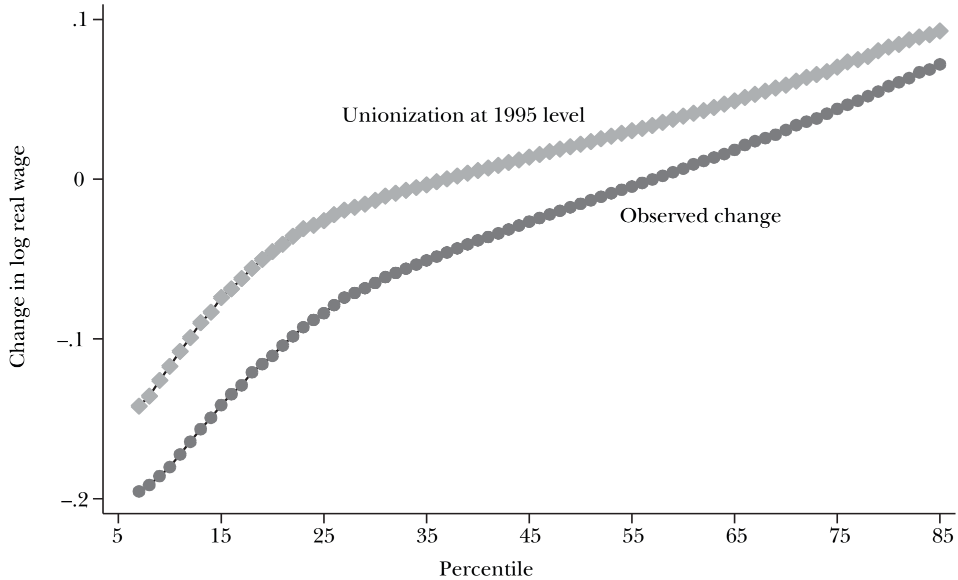
Has this decrease in union coverage rates contributed to a reduction in wage growth and to an increase in inequality? We investigate this question in Figure 5, where we plot the observed changes in log real wages between 1995 and 2008 along the wage distribution. We also plot the counterfactual changes that would have occurred if unionization rates had remained at the same level as in 1995, using the reweighting approach developed in DiNardo, Fortin, and Lemieux (1996), which essentially reweighs wages observed in 2008 with the odds-ratio that a worker with specific observed characteristics has been observed in the 2008-coverage-status in 1995 versus being observed in the 2008-coverage-status in 2008. Notice that this constructed counterfactual exercise is by no means "causal," among other reasons because it ignores general equilibrium effects of de-unionization. The figure suggests that Germany's wages in 2008 would have been higher if union coverage had remained the same as in 1995 throughout the entire wage distribution, but the difference is particularly large at the lower end of the wage distribution.

Works Councils and Opening Clauses

Wage inequality has also increased strongly among employees covered by union contracts, thus suggesting that the German system of industrial relations has allowed for wage adjustments even within the unionized sector. This pattern is illustrated in

[8] See Data Appendix A and Table A3 in Appendix C, available online with the paper at http://e-jep.org.

Figure 5
Observed versus Counterfactual Wage Growth between 1995 and 2008 along the Wage Distribution: The Role of De-unionization

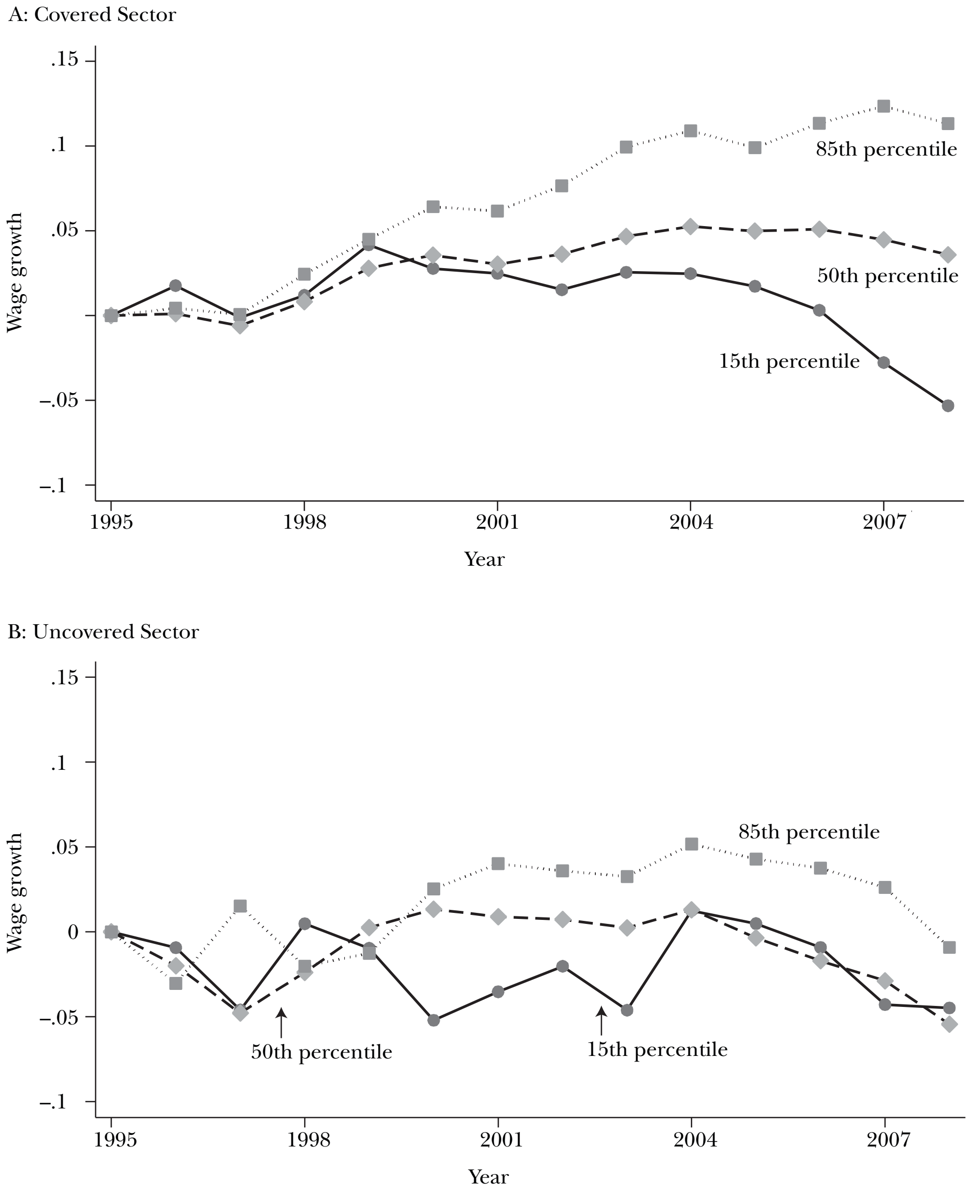


Notes: The figure shows the observed wage growth by percentile between 1995 and 2008, as well as the counterfactual wage growth which would have prevailed if the share of workers covered either by industry-wide or firm-wide agreements had remained at its 1995 level. The counterfactual wage distribution is computed using the reweighting approach developed by DiNardo, Fortin, and Lemieux (1996). Calculations are based in the LIAB.

Figure 6, where we show the evolution of the 15th, 50th, and 85th percentile of the wage distributions, indexed at 0 in 1995, for those covered by a union (panel A) and uncovered by a union (panel B) sectors. The figure shows that wage inequality rose strongly in the covered sector both at the bottom and the top of the wage distribution, while in the uncovered sector it remained basically constant at the bottom of the wage distribution and only increased at the top of the wage distribution. However, notice that due to the indexation the figures hide the larger differentials in wage levels at specific percentiles in the uncovered sector relative to the covered sector: While the 85–50th and 50–15th differentials were on average 0.4 and 0.34 in the covered sector between 1995 and 2008, they were about 0.5 in the uncovered sector. Thus, three factors contributed to the rise in overall inequality during the time period under consideration, namely, the shift of workers from the covered to the uncovered sector (which led, due to the larger differences in wage levels in the uncovered sector, to an increase in lower tail inequality), the increase in inequality in the covered sector, and the increase in inequality at the top of the wage distribution in the uncovered sector.

Figure 6

Indexed Wage Growth of the 15th, 50th, 85th Percentiles, West Germany, by Union Coverage Status, 1995–2008



Notes: Calculations based on LIAB Sample for West German Full-Time Workers between 20 and 60 years of age. The figure shows the indexed (log) real wage growth of the 15th, 50th, and 85th percentiles of the wage distribution, with 1995 as the base year. Nominal wages are deflated using the consumer price index (1995 = 100) provided by the German Federal Statistical Office.

We think that the change in wage inequality in the covered sector is due to the decentralization of wage setting since the beginning of the 1990s, when industry-level collective bargaining came under increasing pressure from employers who demanded more firm-specific and differentiated regulations. Works councils accommodated this decentralization to secure jobs in Germany, which also strengthened their role in the industrial relations. As another response, the trade unions and employers' associations agreed on an increasing number of "opening clauses" in industry-level collective agreements. Opening clauses allow firms to deviate from collectively agreed industry-wide standards. At first, these opening clauses focused on hours of work, but later they also affected wages. Also, the opening clauses were initially only temporary to avoid bankruptcy, but later they were also implemented to ensure competitiveness in more general terms. Firms that use opening clauses negotiate the details concerning pay and working time agreements with the works council. Under German law, firms without a works council cannot use opening clauses, but such firms may instead decide to stop recognizing a union contract. Firms with a works council not covered by a union contract may reach an agreement on wages with the works council.

Brändle, Heinbach, and Meier (2011, Figure 1) report that opening clauses for wages only started to gain importance in 1995 (opening clauses regarding hours of work had existed before 1995). Among industry-wide collective contracts in manufacturing, less than 5 percent involved opening clauses for wages in 1995, but this had risen to about 60 percent in 2004. According to a survey of works councils in 2005, about 75 percent of all firms with collective agreements use opening clauses (Bispinck 2007; Bispinck, Dribbusch, and Schulten 2010).

To summarize, the specific governance structure of the German system of industrial relations allowed for an unprecedented increase in the decentralization of the wage setting process, leading to a decrease in real wages, in particular at the lower end of the wage distribution. This was driven by two main developments: 1) a sharp decline in the share of workers covered by union agreements; and 2) an increase in opening clauses that strengthened the role of firm-based works councils in wage determination relative to trade unions. This argument is consistent with the finding that the rise in firm-level differences in wages contributes strongly to the rise in wage inequality in Germany (Antonczyk, Fitzenberger, and Sommerfeld 2010; Card, Heining, and Kline 2013).

What Led to Greater Flexibility in the German Labor Market?

Why were wage restraints and decreasing real wages at the lower end of the wage distribution in Germany possible after the mid 1990s but not before? After all, German firms have always had the option not to recognize a union contract and to pay wages below the union wage, provided their employees accepted this. Opening clauses had been possible before the mid 1990s. Our answer traces to the major changes in Germany's economy in the early 1990s related to the reunification of Germany and the opening up of the nearby central and eastern European economies.

On one hand, the extraordinary cost of German unification burdened the German economy in an unprecedented way, which is partly responsible for Germany's dismal performance throughout the 1990s and early 2000s. The German Council of Economic Experts (SVR 2004, table 100, p. 644) estimates net transfers of about 900 billion euros from former West Germany to East Germany during the time period 1991 to 2003. The total sum of net transfers corresponds to about half of one year's GDP in Germany during that time period. On the other hand, the opening of central and eastern European countries constituted a unique opportunity for German industry to move production abroad. They offered a stable investment climate, as well as (despite being locked away for several decades behind the Iron Curtain) a long history of trade and interaction with Germany. The structure of industry and education systems, for instance, shared many similarities, which survived the Soviet era. Vocational training plays a key role in the education system, in a way similar to Germany, in countries like Hungary or Poland. German was also widely spoken in parts of Central and Eastern Europe. At the same time, wages in these countries were far lower than in Germany, and working regulations more flexible (for example, Geishecker 2006; Marin 2006). Moving production abroad to these countries took place at a moderate pace: for example, the stock of German foreign direct investment to Poland, Hungary, as well as the Czech and the Slovak Republics amounted to about 1 percent of German GDP in 2000 and about 2.3 percent in 2010 (according to our calculations and data from http://stats.oecd.org/Index.aspx?DatasetCode=FDI_POSITION_PARTNER). However, the possibility that German firms might relocate production to these low-wage countries was very credible, and widely discussed in public (among German media outlets, see the articles in *DIHK* 2003; Mihm and Knop 2004; Hawranek, Hornig, and Jung 2004).

The fiscal burden of German reunification, coupled with an immediately more competitive global environment, made it increasingly costly for German firms to pay high union wages. The new opportunities to move production abroad, while remaining still nearby, changed the power equilibrium between trade unions and employer federations, and forced unions and/or works councils to accept deviations from industry-wide agreements which often resulted in lower wages for workers. In a similar vein, Burda (2000) predicted that the EU-accession of Eastern European countries would foster a reduction of labor market rigidities in the old EU member countries (including Germany). Germany's unions and works councils realized that they had to make concessions in order not to be further marginalized, and the specific characteristics of the German system of industrial institutions allowed the trade unions to adapt to the new economic realities and to make these concessions. As a result, the German labor market appeared to be far more flexible than many would ever have expected.

Why did other continental European countries not react in the same way as Germany? One important reason is that the particularly difficult economic situation in which Germany found itself in the early 1990s was to a large part specific to

Germany, due to the reunification of Germany, which was not felt in other European countries. This was reinforced by Germany's geographic vicinity to the countries of central and eastern Europe, which gave Germany an early taste of the challenges of globalization. This decade of economic stagnation and hardship, when Germany was the "sick man of Europe," prepared the population for accepting agreements for the sake of economic growth, which saw inequality rise dramatically for the first time in the after-war period.

In addition, the system of industrial relations in other continental European countries does not allow for the same inherent opportunities of flexible adaptation as the German system. For example, in countries like France and Italy, union wages are often bargained at the national level and apply to all firms in the economy, regardless of whether the firm explicitly recognizes the union contract. Coverage by union wage contracts has remained remarkably stable at very high levels at about 90 percent in France and 80 percent in Italy during the 1990s and the 2000s (OECD 2004, 2012; Visser 2013). Furthermore, in contrast to Germany, union wage contracts are typically extended to all workers in an industry (OECD 2004, table 3.4, p. 148; Visser 2013, table 4, pp. 96–98). In these and other continental European countries, adding flexibility into collective agreements would require political reforms at the national level. More generally, many of the regulations which are determined by labor contracts in Germany are either legally enforced in other countries (such as the minimum wage in France) or nationally implemented (for example, union agreements extend to all firms in the economy), and therefore require consent on a much higher level (nationally, or even on the political level) to be modified and changed. There is much less scope in these countries for a decentralization of wage setting (and other aspects of working conditions) within their system of industrial relations.

In general, the decentralization of union agreements is certainly being discussed more widely across Europe, but whether or when such changes might occur more widely remains uncertain.

Discussion and Outlook

We have argued that the remarkable transformation of the German economy from the "sick man of Europe" to a lean and highly competitive economy within little more than a decade is rooted in the inherent flexibility of the German system of industrial relations. This system allowed German industry to react appropriately and flexibly over time to the demands of German unification, and the global challenges of a new world economy. However, this intrinsic flexibility became only evident under the extraordinary difficult economic circumstances and the extreme duress in which Germany found itself in the decade after reunification. How does our thesis fit with two other possible explanations for Germany's increased competitiveness: Germany's Hartz labor market reforms of 2003, or the changes brought about by the adoption of the euro?

Germany's government under Gerhard Schröder implemented the so-called "Hartz Reforms" to the labor markets in 2003, which are often credited for spurring Germany's economy (for example, Rinne and Zimmermann 2012, 2013; see Fitzenberger 2009 for a critical assessment of the Hartz Reforms). These reforms were extremely controversial at the time. They reduced and limited the benefits while unemployed, liberalized agency work, reformed "active" labor market policies, and reorganized the Federal Labor Agency, but did not make any institutional changes in the wage setting process.

The Hartz reforms were implemented starting in 2003, hence nearly a decade after the process of wage decentralization and the improvement in competitiveness had begun in Germany. It seems plausible that the changes already underway in Germany's labor markets helped in preparing the political ground for the Hartz reforms. In addition, as the enumeration of the main components of the reforms makes clear, the scale of the reforms is modest enough that they seem unlikely to have triggered the dramatic increase in competitiveness or the enormous drop in German unemployment or to have led Germany's labor market through the deep recession in 2008–2009. Further, while the focus of the reforms was on creating incentives for seeking employment, they did little to support the remarkable wage restraint witnessed since the mid 1990s, which is the key factor in explaining the gain in competitiveness.

We therefore believe that while the Hartz reforms have contributed to the recent decline in long-term unemployment and to the continued increase in wage inequality at the lower end of the wage distribution, they were not central or essential in the process of improving the competitiveness of German industry. Moreover, although one sometimes hears the argument that other continental European countries should muster the political will to adopt their own version of the Hartz reforms, we believe that such a recommendation may be misleading. In our view, the specific governance structure of the German system of industrial relations—activated under extreme duress—is what paved the way for the remarkable decentralization of wage determination from the industry level to the level of the single firm or single worker, and which together with a significant increase in productivity ultimately improved Germany's competitiveness. Whether the political process would have been able to achieve a similar degree of wage decentralization, had the autonomy of wage bargaining not existed in Germany, is doubtful. In our view, the policy recommendation from Germany for the rest of continental Europe should not be the Hartz reforms (the advice given often by policymakers, as in a February 2013 speech by German Chancellor Angela Merkel reported in de Weck 2013), but reforms that would target the system of industrial relations by decentralizing bargaining to the firm level while keeping workers' representatives involved to secure that employees benefit again when economic conditions improve.

Some argue that the adoption of the common European currency is a main factor that has helped Germany to improve competitiveness. Again, we believe that the arrival of the euro may have been a contributing factor, but not the main one.

First, recall that Germany was shifting its labor market institutions and improving its competitiveness during the mid 1990s, and the euro did not start until 1999. Second, within the common currency area, and after 2001, Germany continued to gain competitiveness with respect to its main trading partners such as Italy and Spain. Third, the euro has persistently appreciated against the US dollar, leading to the increase in competitiveness of the United States as we illustrated in Figure 1. It seems unlikely that Germany's deutschemark (if the euro had not been introduced) would have appreciated much more against the dollar than the euro has, at least not before the start of the global financial crisis around 2008 and the ongoing European debt crisis. Finally, it is not clear whether an appreciation of a German currency (which probably would not have taken place before 2008) would have had a dramatic impact on Germany's overall competitiveness at least in the medium-term, because it would also have made imported inputs less costly and it would possibly have fostered even stronger labor market adjustments of the type we have described above. For example, the depreciation of the British pound by nearly 30 percent in 2008–2009 has done little to help UK manufacturing exports.

Of course, the existence of the common euro currency area raises a number of issues for countries within the eurozone. Without the possibility to depreciate national currencies, the only way for countries like France, Italy, and Spain to gain competitiveness relative to other countries of the eurozone is to reduce unit labor costs—that is, by increasing productivity relative to real wages. Whether these countries will succeed in this endeavor remains an open question. The more centralized and legally anchored nature of their labor market institutions, in comparison to Germany, does put them at a disadvantage in making such an adaptation. Boeri (2011) provides an assessment of the political economy of labor market reforms with a particular focus on countries of southern Europe. He argues that the political process often allows only for two-tier reforms (affecting only a subset of all employees) instead of complete reforms, which may not result in an increase of competitiveness

The rise in inequality in Germany has led to an intensive debate about its social consequences, and its effect on poverty and social justice. For example, recent negotiations between employers and employee associations in Germany suggest that future wage settlements will try to make up for the loss in real wages many workers experienced in recent decades. It is also likely that certain aspects of labor and wage regulations will in the future be "put in legislative stone." As one example, the new coalition government in Germany will introduce a nationally legislated minimum wage. Thus, the possibility for Germany to rely on its system of industrial relations to improve its competitive position by having a decentralized decision making process may be cut back, and this may restrict Germany's ability to react in similar ways to future economic challenges. If that occurs, then future gains in German competitiveness will need to be accomplished rather through increases in productivity that outstrip wage increases. This pattern may help to bring convergence in the competitiveness of the countries in the eurozone.

■ *Christian Dustmann is grateful for insightful and stimulating discussions with Wendy Carlin. We thank Wendy for comments on earlier drafts. We also benefited from excellent research assistance by Marina Furdas, Stefanie Licklederer, Olga Orlanski, and Florian Weiss. Bernd Fitzenberger dedicates his contribution for this paper to John Pencavel, his fabulous teacher in labor economics.*

References

Addison, John T., Claus Schnabel, and Joaquim Wagner. 1997. "On the Determinants of Mandatory Works Councils in Germany." *Industrial Relations* 36(4): 419–45.

Antonczyk, Dirk, Bernd Fitzenberger, and Katrin Sommerfeld. 2010. "Rising Wage Inequality, the Decline of Collective Bargaining, and the Gender Wage Gap." *Labour Economics* 17(5): 835–47.

Beckmann, Michael, Sylvia Föhr, and Matthias Kräkel. 2010. "Rent Seeking, Employment Security and Works Councils: Theory and Evidence for Germany." *Schmalenbach Business Review* 62(1): 2–40.

Bispinck, Reinhard. 2007. "Löhne, Tarifverhandlungen und Tarifsystem in Deutschland 1995–2005." WSI Discussion Paper No. 150, Wirtschafts- und Sozialwissenschaftliches Institut, Düsseldorf.

Bispinck, Reinhard, Heiner Dribbusch, and Thorsten Schulten. 2010. "German Collective Bargaining in a European Perspective: Continuous Erosion or Re-Stabilisation of Multi-Employer Agreements?" WSI Discussion Paper No. 171, Wirtschafts- und Sozialwissenschaftliches Institut, Düsseldorf.

Boeri, Tito. 2011. "Institutional Reforms and Dualism in European Labor Markets." Chap. 13 in Ashenfelter, Orley, and David Card, eds., *Handbook of Labor Economics*, Volume 4A and 4B. North Holland, Elsevier.

Brändle, T., W. D. Heinbach, and M. F. Meier. 2011. "Tarifliche Öffnung in Deutschland: Ausmaß, Determinanten, Auswirkungen." *Journal for Labour Market Research* 44(1–2): 163–72.

Burda, Michael. 2000. "Mehr Arbeitslose—der Preis für die Osterweiterung? Zur Auswirkung der EU-Osterweiterung auf die europäischen Arbeitsmärkte im Osten und Westen [More unemployment—the price of eastwards expansion? The impact of the eastwards expansion of the EU on European labor markets in the East and West]." *Duncker & Humblot, Serientitel: Schriften des Vereins für Socialpolitik Neue Folge* 274, 79–102.

Burda, Michael, and Jennifer Hunt. 2011. "What Explains the German Labor Market Miracle in the Great Recession?" *Brookings Papers on Economic Activity* 42(1): 273–335.

Card, David, Joerg Heining, and Patrick Kline. 2013. "Workplace Heterogeneity and the Rise of German Wage Inequality." *Quarterly Journal of Economics* 128(3): 967–1015.

Carlin, Wendy. 2012. "Real Exchange Rate Adjustment, Wage-setting Institutions, and Fiscal Stabilization Policy: Lessons of the Eurozone's First Decade." *CESifo Economic Studies* 59(3): 489–519.

Carlin, Wendy, and David Soskice. 2008. "Reforms, Macroeconomic Policy and Economic Performance in Germany." In *Economic Policy Proposals for Germany and Europe,* edited by Ronald Schettkat, and Jochem Langkau, 72–118. Routledge.

Carlin, Wendy, and David Soskice. 2009. "German Economic Performance: Disentangling the Role of Supply-side Reforms, Macroeconomic Policy and Coordinated Economy Institutions." *Socio-Economic Review* 7(1): 67–99.

Dauth, Wolfgang, Sebastian Findeisen, and Jens Suedekum. 2012. "The Rise of the East and the Far East: German Labor Markets and Trade Integration." IZA Discussion Paper No. 6685.

de Weck, Joseph. 2013. "Merkel Cites East

German Lessons for EU's Problem States. *Bloomberg.com.* February 19. http://www.bloomberg.com/news/2013-02-18/merkel-cites-east-german-lessons-for-crisis-wracked-euro-states.html.

DIHK [Deutscher Industrie- und Handelskammertag]. 2003. "Produktionsverlagerung als Element der Globalisierungsstrategie von Unternehmen: Ergebnisse einer Unternehmensbefragung Mai 2003." ["Offshoring of Production as Part of the Globalization Strategy of Firms: Results of a Firm Survey in May 2003, Berlin."] http://www.dihk.de/ressourcen/downloads/produktionsverlagerung.pdf.

DiNardo, John, Nicole Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach." *Econometrica* 64(5): 1001–1044.

Dorner, Matthias, Jörg Heining, Peter Jacobebbinghaus, and Stephan Seth. 2010. "Sample of Integrated Labour Market Biographies (SIAB) 1975–2008." FDZ-Datenreport 01/2010.

Dustmann, Christian, and Uta Schönberg. 2009. "Training and Union Wages." *Review of Economics and Statistics* 91(2): 363–76.

Dustmann, Christian, Johannes Ludsteck, and Uta Schönberg. 2009. "Revisiting the German Wage Structure." *Quarterly Journal of Economics* 124(2): 843–81.

Economist. 2004. "Germany on the Mend." November 17. http://www.economist.com/node/3352024.

Egger, Harmut, and Peter Egger. 2003. "Outsourcing and Skill Specific Employment in a Small Economy: Austria after the Fall of the Iron Curtain." *Oxford Economic Papers* 55(4): 625–43.

Fitzenberger, Bernd. 2009. "Nach der Reform ist vor der Reform? Eine arbeitsökonomische Analyse ausgewählter Aspekte der Hartz-Reformen." In *Reformen für Deutschland—Die wichtigsten Handlungsfelder aus ökonomischer Sicht,*" edited by Günther Schulze pp. 21–48. Schäffer-Poeschel Verlag.

Fitzenberger, Bernd, Karsten Kohn, and Alexander C. Lembcke. 2013. "Union Density and Varieties of Coverage: The Anatomy of Union Wage Effects in Germany." *Industrial Labor Relations Review* 66(1): 169–97.

Fitzenberger, Bernd, Karsten Kohn, and Qingwei Wang. 2011. "The Erosion of Union Membership in Germany: Determinants, Densities, Decompositions." *Journal of Population Economics* 24(1): 141–65.

Gartner, Hermann. 2005. "The Imputation of Wages above the Contribution Limit with the German IAB Employment Sample." FDZ-Methodenreport 02/2005.

Geishecker, Ingo. 2006. "Does Outsourcing to Central and Eastern Europe really threaten Manual Workers' Jobs in Germany?" *World Economy* 29(5): 559–83.

Greenhouse, Steven. 2013. "VW and Its Workers Explore a Union at a Tennessee Plant." *New York Times,* September 6. http://www.nytimes.com/2013/09/07/business/vw-and-auto-workers-explore-union-at-tennessee-plant.html.

Hassel, Anke. 1999. "The Erosion of the German System of Industrial Relations." *British Journal of Industrial Relations* 37(3): 483–505.

Hassel, Anke, and Britta Rehder. 2001. "Institutional Change in the German Wage Bargaining System—The Role of Big Companies." MPIfG Working Paper 01/9.

Hawranek, Dietmar, Frank Hornig, and Alexander Jung. 2004. "Bye-bye 'Made in Germany.'" *Der Spiegel* 44: 94–109.

Jacobebbinghaus, Peter, and Stephan Seth. 2010. "Linked Employer-Employee Data from the IAB: LIAB Cross-sectional (Model 2, 1993-2008)." FDZ-Datenreport 05/2010.

Lesch, Hagen. 2009. "Erfassung und Entwicklung von Streiks in OECD-Ländern." *IW Trends,* 1/2009, Institut der deutschen Wirtschaft, Koln.

Marin, Dalia. 2006. "A New International Division of Labor in Europe: Outsourcing and Offshoring to Eastern Europe." *Journal of the European Economic Association* 4(2/3): 612–22.

Mihm, Andreas, and Carsten Knop. 2004. "Outsourcing: Bis ich mit der Gewerkschaft klar bin, hat die Konkurrenz schon geliefert [Outsourcing: Until I have reached an agreement with the union, the competitor has already delivered]." *FAZ [Frankfurter Allgemeine Zeitung],* March 21. http://www.faz.net/aktuell/wirtschaft/outsourcing-bis-ich-mit-der-gewerkschaft-klar-bin-hat-die-konkurrenz-schon-geliefert-1144367.html.

Möller, Joachim. 2010. "The German Labor Market Response in the World Recession—De-mystifying a Miracle." *Journal for Labour Market Research* 42(4): 325–36.

OECD. 2004, 2007, 2012. *Employment Outlook.*

Pilat, Dirk, Agnès Cimper, Karsten Olsen, and Colin Webb. 2006. "The Changing Nature of Manufacturing in OECD Economies." OECD, STI Working Paper 2006/9.

Rinne, Ulf, and Klaus F. Zimmermann. 2012. "Another Economic Miracle? The German Labor Market and the Great Recession." *IZA Journal of Labor Policy* 1/3.

Rinne, Ulf, and Klaus F. Zimmermann. 2013. "Is Germany the North Star of Labor Market Policy?" *IMF Economic Review* 61(4): 702–729.

Sinn, Hans-Werner. 2006. "The Pathological Export Boom and the Bazaar Effect: How to Solve the German Puzzle." *World Economy* 29(9): 1157–75.

Spermann, Alexander. 2012. "The New Role of Temporary Agency Work in Germany." In *Die neue Rolle der Zeitarbeit in Deutschland, Beiträge zur Flexibilisierung* edited by Richarda B. Bouncken, Lutz Bellmann, and Manfred Bornewasser, Volume 3, 203–224. München and Mering.

StaBu [German Federal Statistical Office]. 2013. Verdienste und Arbeitskosten–Tarifbindung in Deutschland 2010. Statistisches Bundesamt, Wiesbaden. https://www.destatis.de/DE/Publikationen/Thematisch/VerdiensteArbeitskosten/Tarifverdienste/Tarifbindung.html.

Streeck, Wolfgang. 1997: "German Capitalism: Does It Exist? Can It Survive?" *New Political Economy* 2(2): 237–56.

SVR [German Council of Economic Experts]. 2004. "Erfolge im Ausland—Herausforderungen im Inland." Annual Report to the Government, Wiesbaden. http://www.sachverstaendigenrat-wirtschaft.de/fruehere_jahresgutachten.html.

Visser, Jelle. 2013. "Wage Bargaining Institutions from Crisis to Crisis." Economic Papers 488, European Commission. April. http://ec.europa.eu/economy_finance/publications/economic_paper/2013/ecp488_en.htm.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 189–208

When Ideas Trump Interests: Preferences, Worldviews, and Policy Innovations

Dani Rodrik

Ideas are strangely absent from modern models of political economy. In most prevailing theories of policy choice, the dominant role is instead played by "vested interests"—elites, lobbies, and rent-seeking groups which get their way at the expense of the general public. Economists, political scientists, and other social scientists appeal to the power of special interests to explain key puzzles in regulation, international trade, economic growth and development, and many other fields. Why does regulation favor incumbent firms at the expense of consumers or potential entrants? Because bureaucracies can be captured by well-organized "insiders" who can shape regulation in their favor (Stigler 1971; Peltzman 1976; Laffont and Tirole 1991). Why are trade restrictions so rampant despite the well-known gains from trade? Because import tariffs and quotas redistribute incomes to politically powerful business groups and lobbies (Krueger 1974; Grossman and Helpman 1994; Rodrik 1995). Why do political elites not favor growth-promoting policies and institutions? Because growth-suppressing policies, such as weak property rights, excessive regulation, or overvalued currencies provide these elites with access to rents that would disappear otherwise (Bates 1981; Acemoglu and Robinson 2006, 2012). Insights from political-economy models in each of these fields exert a strong influence on the way economists think of societal outcomes and the operation of the political system.

Any model of political economy in which organized interests do not figure prominently is likely to remain vacuous and incomplete. But it does not follow from this that interests are the ultimate determinant of political outcomes. Here

■ *Dani Rodrik is the Albert O. Hirschman Professor in the School of Social Science, Institute for Advanced Study, Princeton, New Jersey. His email address is drodrik@ias.edu.*

http://dx.doi.org/10.1257/jep.28.1.189 doi=10.1257/jep.28.1.189

I will challenge the notion that there is a well-defined mapping from "interests" to outcomes. This mapping depends on many unstated assumptions about the *ideas* that political agents have about: 1) what they are maximizing, 2) how the world works, and 3) the set of tools they have at their disposal to further their interests. Importantly, these ideas are subject to both manipulation and innovation, making them part of the political game. There is, in fact, a direct parallel, as I will show, between inventive activity in technology, which economists now routinely make endogenous in their models, and investment in persuasion and policy innovation in the political arena. Once their fluid nature is recognized, vested interests become much less determining and the space of possible outcomes much wider.

While hidden assumptions play a role in all economic models, I will argue that the failure to recognize the role of ideas in shaping interests (and their pursuit) has especially serious implications in political economy. Taking ideas into account allows us to provide a more convincing account of both stasis and change in political-economic life. It provides a way of bridging the sharp divide between policy analysis (what should be done) with political economy (what actually happens). It also yields an explanation for a common anomaly in the real world—the fact that, in practice, many reforms turn out to benefit the elites who previously blocked them.

I begin by providing a taxonomy of the ways in which ideas make their way, more often than not implicitly, into established models of political economy. Next, I focus specifically on models professing to explain economic inefficiency and argue that outcomes in such models are determined as much by the ideas that elites are presumed to have on feasible strategies as by vested interests themselves. A corollary is that new ideas about policy—or policy entrepreneurship—can exert an independent effect on equilibrium outcomes even in the absence of changes in the configuration of political power. I conclude by discussing the sources of new ideas.

Making the Role of Ideas Explicit

Every rational-choice model is built on the purposive behavior of individual decision makers. Typically, behavior is determined by assuming individuals solve a well-defined optimization problem. At least three components must be specified in such an optimization exercise: 1) an objective function (like a consumer utility function); 2) a set of constraints (like a budget constraint); and 3) a set of choice variables (like the quantities of consumption to be chosen). Political economy models in the rational-choice mold translate this framework into the political arena. Political agents—voters, lobbies, elites, congressmen—are represented as rational individuals who solve explicit optimization problems. This means: 1) they maximize a utility function defined over consumption, rents, or political benefits; 2) they operate within constraints imposed by the rules of the game, both economic and political; and 3) they choose a set of actions—which in various models may include votes, political contributions, rebellion, and suppression—that maximize their objection function given the constraints. For example, business lobbies decide how much they should make in political contributions in return for tariff protection, taking into account that politicians value societal welfare alongside the contributions (Grossman and Helpman 1994). Or a dictator decides whether to develop his economy so as to maximize his intertemporal stream of rents, taking into account that his decision affects both economic and political outcomes, including his longevity in power (Olson 1993; Acemoglu and Robinson 2006).[1]

Ideas enter this problem in several distinct ways that are rarely recognized. In fact, each of the three components of the optimization problem—preferences, constraints, and choice variables—rely on an implicit set of ideas. I discuss these in turn below. I emphasize that I am not contesting rationality or the utility of the basic optimizing framework in the political arena. My goal is to explore the role of ideas in shaping how interests are defined and pursued.

Preferences: Who Are We?

Self-interest presumes the idea of a "self"—that is, a conception of who I am and what my purpose is. In many economic applications, the maximand is clear. It is reasonable to assume households want to maximize their consumer surplus and producers their profits, although these assumptions are not always entirely uncontroversial. In the political sphere, the choice of what is to be maximized is much less evident: depending on context, honor, glory, reputation, respect, income, power, durability in office, and "good of the country" are all plausible. As Elster (2000) has written in a critique of political economy frameworks attempting to explain historical political developments, the nobility in 17th-century France may have been interested in honor and glory as much as in material benefits. Much human behavior is driven by abstract ideals, sacred values, or conceptions of loyalty that cannot be reduced to economic ends. Studies by anthropologists and psychologists suggest "humans will kill and die not only to protect their own lives or defend kin and kith, but for an idea—the moral conception they form of themselves, of 'who we are'" (Atran and Ginges 2012, p. 855)—a point that should not be controversial in an age of suicide bombings.

[1] More formally, we can express the canonical political-economy problem as follows. Let $\boldsymbol{a}$ denote the vector of choice variables, $\boldsymbol{c}$ the gross private consumption vector, and α the share of the economy's surplus (or gross private consumption) that political agents can extract. Then the optimization problem that each agent solves takes the following form:

$$\max_{\{\boldsymbol{a}\}} U(\alpha \boldsymbol{c}, .) \quad \text{s.t.} \quad \begin{aligned} \varphi(\boldsymbol{a}, \boldsymbol{c}; \boldsymbol{\theta}) &\leq 0 \\ \alpha &= g(\boldsymbol{a}, \boldsymbol{c}; \boldsymbol{\theta}) \end{aligned}$$

where $\boldsymbol{\theta}$ is a vector of exogenous parameters. The agent's utility depends, at least in part (and not necessarily exclusively), on the surplus that the agent can extract. The constraints in this optimization exercise come in two forms, specifying the economic and political model, respectively. The first relationship $\varphi(.)$ describes the economic equilibrium, which is determined jointly by actions and exogenous parameters. The second relationship describes the political mechanism, represented by $g(.)$, that specifies how political actions, along with other endogenous and exogenous variables, translate into rent extraction. While time subscripts are not included, it should be clear how this framework can be extended to a dynamic setting.

How we evaluate different social states and judge whether they advance our "interests" depends crucially on how we define ourselves. We might view ourselves as a member of a social class ("middle class"), ethnic group ("white majority"), religion ("evangelical"), nation ("global citizen"), demographic cohort ("baby boomer"), profession ("educator"), or a myriad of other possible identities. As Sen (2006) has suggested, we might even combine all these identities in varying degrees. In political science, a well-established line of research has long held that interests of political actors are socially constructed rather than determined by well-defined material facts. In this "constructivist" tradition, conceptions of interests arise endogenously from norms, ideologies, and causal beliefs (Wendt 1999; Ruggie 1998; Blyth 2002; Hay 2011). Interests, in fact, are "one form of idea" (Béland and Cox 2011, p. 10). In international law, a parallel discussion pits "legal realists," who argue that behavior among states is determined exclusively or largely by national interests, against scholars who see a significant role for norms of justice or law (Howse 2013; Goldsmith and Posner 2005).

Economists have rarely ventured that far, but the role of ideas in determining preferences has crept into various strands of research in economics. For example, the partisan-politics literature in macroeconomics endows political parties with explicit ideologies, typically represented as different preference weights on inflation versus unemployment (Alesina and Rosenthal 1995). These differences in preferences are typically imposed exogenously, with little explanation. More recently there has been some work, both at the macro and micro level, that looks at the endogenous determination of ideologies: see, for example, Alesina, Cozzi, and Mantovan (2012), Della Vigna and Kaplan (2007), and Yanagizawa-Drott (2012) on the formation of political preferences through exposure to societal outcomes, media, or early childhood experiences.

The work of Akerlof and Kranton (2000; 2005 in this journal) on the economics of identity is particularly relevant here. Akerlof and Kranton consider models where individuals associate themselves with specific social categories and their desired behavior derives from the attributes of those categories. Workers, for example, may acquire identities that moderate the incentive compatibility constraint vis-à-vis their employers (leading them to behave in greater conformity with the objectives of firms). In turn, employers may seek to alter such identities to enhance workplace performance. Such models could explain a range of "anomalous" political action, including voting against immediate material interests. But the possible application of this approach to political phenomena has not yet received much attention.

In each of these frameworks, ideas play a crucial role. Instead of being epiphenomenal, they determine preferences directly, and therefore shape patterns of political behavior. This perspective requires social scientists to engage with the questions of where ideas come from and how they are spread and internalized.

The Model: How Does the World Work?

Policymakers operate under certain working assumptions about how the world works. Their worldviews shape their perception of the consequences of their and others' actions in both economic and political domains. These ideas may fall on either side of some of the biggest controversies in the history of economic thought: Does the economy work better under laissez-faire or planning? Are economic growth and development more rapid under free trade or under protection? Does macroeconomic stability require Keynesian countercyclical policies or Hayekian nonintervention? Each of these positions presumes a particular model of how the economy works and therefore has different implications for political behavior. In recent decades, a succession of economic ideas (think of Keynesianism, monetarism, rational expectations, the "Washington consensus") have served to change elites' understanding of "economic reality" and thereby altered the political equilibrium.

A rent-extracting autocrat is likely to tax subjects to the hilt when he believes they have little choice but to pay up; the same autocrat will be more restrained when he believes subjects can evade the tax or mount an effective opposition. "Elasticity pessimism," the belief that economic activities tend to be unresponsive to prices, lay behind the widespread consensus in support of the *dirigiste* economic development policies of the 1950s and 1960s. Mehta and Walton (2012, pp. 17–18) describe the Nehruvian cognitive map that shaped India's development path in the decades following independence: the need for a big investment push, suspicions over the private sector, emphasis on the leading role of capital goods, and export pessimism (the fear that expansion of exports will run against serious limits) all derived from ideas about how a market system worked (or failed to do so). As research demonstrated that the poor were as responsive to price incentives as the rich, policies in the developing world began to move in a more market-oriented direction (Schultz 1964; Krueger 1997).

Economic research is all about sharpening our ideas about the "right model." Yet actors in political economy models live in worlds where these questions have been effectively resolved. They believe that they know how the world works, if not precisely, at least probabilistically. Rational expectations mean agents base their expectations, and hence behavior, on how the world really works. Even if they disagreed for a time, adherents of this view argue, agents would eventually reach agreement on the "right model" as events and reactions to policy choices unfold. In practice, however, people often downplay evidence that seems inconsistent with their model of the world. Anomalous outcomes are dismissed as a fluke or as the result of insufficiently vigorous application of their preferred policy. Even with rational Bayesian updating, a common signal may exacerbate posteriors—increase polarization—when priors differ (Dixit and Weibull 2007).[2] And divergences in beliefs need not disappear even asymptotically when there is disagreement over the interpretation (that is, informativeness) of the signals received (Acemoglu, Chernozhukov, and Yildiz 2009).

[2] An example provided by Dixit and Weibull (2007, p. 7353): Observing a sudden rise in unemployment, a Keynesian might strengthen his belief that monetary policy is too tight, while a monetarist might infer monetary policy is even looser than he thought and is discouraging employment creation by generating expectations of future inflation.

A more realistic representation may be that cognitive and other limitations force political agents to live in a world of Knightian uncertainty with respect to their understanding of causal relationships (see, for example, Denzau and North 1994; Blyth 2011). Their view of the world could be wrong and could remain so even in the face of new evidence if that evidence is just used to confirm past beliefs. Conversely new information may present realities previously not considered. For instance, voters may discover that an office holder has a long criminal track record, a possibility that may not have entered their calculus earlier. An interesting new empirical literature has begun to document how the provision of such information can influence voters' behavior (Pande 2011).

Consider the experience of the global economic and financial crisis in recent years and the extent to which it has altered beliefs. Many observers, such as Johnson and Kwak (2010), have argued that the policies that produced the crisis were the result of powerful banking and financial interests getting their way, which seems like a straightforward application of the theory of special interests. Still, without the wave of ideas "in the air" that favored financial liberalization and self-regulation and emphasized the impossibility (or undesirability) of government regulation, these vested interests would not have gotten nearly as much traction as they did. After all, powerful interests rarely get their way in a democracy by nakedly arguing for their own self-interest. Instead, they seek legitimacy for their arguments by saying these policies are in the public interest. The argument in favor of financial deregulation was not that it was good for Wall Street, but that it was good for Main Street. Other observers have argued that the financial crisis was a result of excessive government intervention to support housing markets, especially for lower-income borrowers. These arguments were also grounded on certain ideas—about the social value of homeownership and the inattentiveness of the financial sector to those with lower incomes. Again, ideas apparently shaped politicians' views of how the world works—and therefore their interest in acting in ways that precipitated the crisis. Finally, although all parties have observed the same Great Recession taking place, relatively few have altered their fundamental beliefs about whether the financial sector is over- or underregulated as a result.

The Strategy Space: What Can Political Actors Do?

Much politics is about strategy: setting the agenda, making alliances, issuing promises (or threats), expanding or restricting the menu of options, building or spending political capital. Regardless of whether they aim to enrich themselves or further a broader set of interests, politicians must continually ask: "What can be done?" Typically, political economy models restrict the strategy space in arbitrary ways: they overstructure the political game by limiting, for example, the policy choices on the menu. This limitation is usually put in place for tractability, but it is not harmless. New ideas about what can be done—innovative policies—can unlock what otherwise might seem like the iron grip of vested interests. The history of economics is replete with ways ideas can shape politics (see Leighton and López, 2013, for a recent useful discussion).

The formal parallelism between behavior in the political arena and consumer behavior in the marketplace is least useful here. Consumers in a market have well-defined choices to make: how much to consume of each available good, given prices and their budget constraint. The standard utility maximization problem does not do great injustice to their strategy space. By contrast, political agents design their own strategy space. The available instruments are up for grabs and limited only by their political imagination.

In what follows, I will focus on this third aspect of ideas—"ideas as policy innovation"—both because it has received relatively little attention in the literature and because it ties in directly with the kind of policy analysis in which economists often engage. I will illustrate the importance of such ideas in an important subclass of political economy models—those that focus on explaining inefficient policy outcomes.

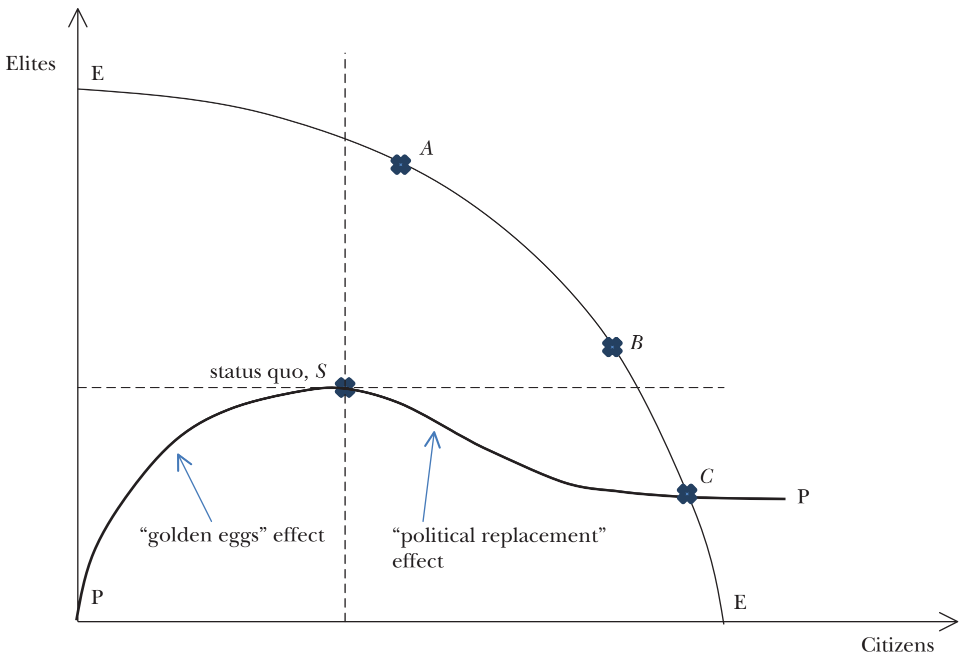
The Political Economy of Inefficient Policy Choice

Explaining why political systems end up making economically inefficient policy choices is a major preoccupation of the political economy literature; it can even be said to be its original raison d'être. Taking ideas explicitly into account, especially in terms of policy innovation, can shed substantial new light on existing approaches. I will show how widely accepted results often rest on arbitrary restrictions on political actors' strategy space. More positively, I argue that an ideas-based perspective allows us to better understand the circumstances under which political systems are, in fact, able to move towards more efficient outcomes.

Political economy models that generate equilibrium policy inefficiency rest on several building blocks: 1) interests determine policy preferences; 2) political power determines whose interests matter (more); and 3) political institutions, or "rules of the game," determine the specific political equilibrium among many. However, while these three postulates can explain redistribution from less-powerful to more-powerful groups, they do not explain inefficiency. Consider Figure 1, which, for the sake of concreteness, divides society into two groups: elites—defined as the group that has both veto and agenda-setting power—and politically powerless citizens. (Ignore the curve PP for now.) At best, the three postulates can help us understand why we are at point like *A* instead of *B*. They cannot explain why we often end up *inside* the economic possibilities frontier at a point like *S*. The equilibrium at *S* violates Pareto-efficiency; it would be possible to improve the incomes of elites without harming citizens. If all elites want is to extract income from society and they are powerful enough to get their way, why do they have to do it by generating inefficiency?

To generate inefficiency, political economy models must add one or both of two additional features, each of which restricts the feasible set of elite actions: 4) more efficient redistributive mechanisms, including lump-sum transfers, are unavailable; and 5) political power is endogenous, and outcomes that move the economy closer to the efficient frontier may reduce the power of elites. The first of these additional assumptions rules out myriads of compensatory policies that can

Figure 1
Redistribution, Inefficiency, and the Political Transformation Frontier



Notes: EE is the economic possibilities frontier. The curve PP is a "political transformation frontier," defined as the set of maximal economic outcomes achievable by those currently in power, the elites. The rising part of the PP curve represents the "goose that lays the golden eggs effect"; along this part of the curve, elites benefit from the enrichment of citizens as a whole, so there is no tradeoff in this region between economic efficiency and political redistribution. The declining part of the curve represents what Acemoglu and Robinson (2006) call "the replacement effect": elites avoid efficient policies because they think these policies will undercut their power and lead to their replacement. When elites act strategically they will select the maximum point on the political transformation frontier. The equilibrium outcome is labeled "status quo," which is far inside the economic transformation frontier.

move the equilibrium from *S* closer to *A*. For example, most forms of economic liberalization, such as removing a tax or an import tariff, are not Pareto-optimal unless accompanied by compensation. Ruling out compensation of elites that lose out from liberalization is an easy way to generate inefficient outcomes. In international economics, for example, policies that restrict trade are third-best for redistributive purposes. So the typical practice in the literature on the political economy of trade policy is to assume away not only lump-sum transfers but also producer subsidies in order to create pressure, within the model, for policies that restrict trade (Rodrik 1995).

A similar strategy is at play in explaining why productivity-enhancing privatizations of state-owned enterprises do not take place: the insiders (workers, managers) would lose out for sure if there is no possibility to compensate them for losses they would experience in a privatized firm. Such restrictions are often rationalized by appealing to an inability to commit to compensation,[3] even though innovative strategies can be often found to get around this constraint, as I will show below.

The other argument, which posits a feedback from economics back to politics, opens up the possibility that elites avoid efficient policies for fear that such policies may undercut their political power and hence their ability to determine future policies. Under this scenario, the only feasible move to the efficient frontier involves traveling to a point such as *C* in Figure 1, where the elites would be worse off.

This type of argument has been invoked in the work of Acemoglu and Robinson (2006, see also Acemoglu 2003) to explain, for example, why some states have blocked policies that would foster industrialization and economic growth during the 19th century in Europe. Because economic growth uproots people from their traditional rural base and facilitates collective political action, it can destabilize entrenched elites. Forward-looking elites will prefer to ensure their power is not challenged, even if that means more inefficiency and less growth.

A different mechanism that produces a similar result is the dynamic inconsistency of compensation policies. In the model we develop in Fernandez and Rodrik (1991), reform is impeded by individual-specific uncertainty combined with the pattern of information revelation over time when reform is implemented. Beneficiaries of today's inefficient policies cannot be made to accept a reform-now-and-compensation-later package because they know reform will reveal the identity of a large group of winners and shift future political power to those winners. After the reform, the beneficiaries will have no need (or incentive) to carry out their promise of compensation.

While the claim that elites block enhanced economic opportunities so that they can maintain their own power seems to make sense, it too may imply an unreasonable restriction on feasible strategies. In particular, it denies elites the creativity to devise strategies that would allow them to take advantage of enhanced economic opportunities without losing power. Unless we posit a highly mechanical relationship between economics and politics, such strategies will exist in general. The restriction on feasible strategies is often relaxed in practice, as we shall see, by innovation in the area of "political ideas." I use the term "innovation" deliberately, as there is a useful analogy here with technological innovation. Just as we think of technological ideas as those that relax resource constraints, we can think of political ideas as those that relax political constraints, enabling those in power to make themselves (and possibly the rest of society) better off without undermining their political power. To develop this point, I introduce the heuristic device of a "political transformation frontier," akin to the economic transformation constraint.

[3] A subtle argument against lump-sum transfers specifically, mentioned by Avinash Dixit in his comments on this paper, is that even atomistic agents will have the incentive to misrepresent their "type" (Hammond 1979). For a good discussion of the political transaction costs that block efficiency, see Dixit (1996, especially chapter 2).

The Political Transformation Frontier and Relaxing Political Constraints

I define the "political transformation frontier" as the set of maximal economic outcomes achievable by those currently in power, the elites. It depicts the maximum level of rents that elites can extract in political-economic equilibrium, taking into account, in particular, the endogeneity of their political power. Return to Figure 1, which depicts a production possibility frontier that shows the tradeoffs between resources consumed by political elites and those consumed by other citizens. The political transformation frontier is depicted as the *PP* schedule in Figure 1. The rising part of the *PP* curve represents what we may call "the goose that lays the golden eggs effect"; along this part of the curve, elites benefit from the enrichment of citizens as a whole and, therefore, there is no tradeoff in this region between economic efficiency and political redistribution. Elites behave here like Mancur Olson's (1993) "stationary bandits," brigands who stay long enough with the farmers from whom they steal that they have an incentive to encourage a degree of prosperity. The declining part of the curve, in turn, represents what Acemoglu and Robinson (2006) call "the replacement effect": enrichment of citizens now comes at the expense of the ability of elites to cling to power and, therefore, their ability to extract rents. When elites act strategically (as Stackelberg leaders) they will select the maximum point on the political transformation frontier. The equilibrium outcome is labeled "status quo," which is far inside the economic transformation frontier.

As the figure is drawn, politics is the binding constraint on this economy. The allocation of resources and overall economic performance are determined not by technological or resource constraints, but by the location and shape of the political transformation frontier *PP*. In turn, the location and shape of the political transformation frontier depends on the set of feasible elite actions—which is determined by the *ideas* that elites have about the range of strategies available to them, including policies, actions, coalitions and so on. Good political/policy ideas relax the political constraint, moving the political transformation frontier (*PP*) outward. Just as technological innovations relax the conventional resource constraint, political innovation relaxes the political transformation frontier. Economists recognize the importance of technological innovation and have made it the centerpiece of their models of long-term growth. In political economy models, by contrast, the working assumption is that there is no room for discovery. Many political innovations are likely to remain ephemeral, inconsequential, or soon forgotten. But some, as with general-purpose technology, may prove substantial and durable: for examples, think of political parties, independence of judiciary, or indeed democracy.

Technological change need not make everyone better off. Similarly, policy innovation could leave the non-elites worse off. (An outward shift of *PP* in Figure 1 may move the new equilibrium to the northwest of *S*.) Some ideas could be bad from the standpoint of society at large and yet gain currency: imagine elites successfully persuading non-elites that they should work harder in this life so they can have redemption in the next, or (closer to home) that an extremely low rate of capital taxation actually benefits them.

Although the idea of a political transformation frontier rarely receives much attention in political economy, practical economists and policymakers do spend considerable effort generating new policy ideas that seek, not always successfully, to work around political constraints. Perhaps the most telling examples of how political innovation can move the political transformation frontier come from the real world. Here are some illustrations that made elites and non-elites both better off.

Achieving Industrialization without Losing Power

Acemoglu and Robinson (2006) discuss how several governments evaded the trap of blocking industrialization for fear of being replaced, and in doing so, they illustrate the feasibility of strategies that shift the political transformation frontier outward. Consider why the Japanese elite decided to spur industrialization and economic development after the Meiji restoration circa 1868. Acemoglu and Robinson note that "the drive for modernization in Japan took a special form, strengthening the centralized government and increasing the entrenchment of bureaucratic elites. In terms of our model, this can be viewed as a strategy to industrialize while also minimizing the probability of replacement . . ." They add that this was similar to what happened in Britain and Germany where "the nonindustrial elites maintained their political power despite the process of industrialization" (p. 128). In Britain, "by adopting a strategy of gradual concessions, [elites] were able to control the political equilibrium and maintain power for at least a century following the onset of the political impact of industrialization . . . [and] the long history of Britain as a trading nation and mercantile power meant that many aristocrats had relatively diversified wealth . . ." (p. 126). In Germany, "the Junkers forged the coalition of 'Iron and Rye' with the rising industrial class to secure their economic interests" (p. 126).

What stands out in each of these explanations is the purposive strategy pursued by elites to mitigate their concern over being replaced as a result of economic change. State-directed industrialization, gradual concessions to the rising industrial classes, diversification into commerce and industry, alliance with industrial interests, and similar choices ensured elites could benefit from industrialization while retaining much of their power. The existence of such strategies raises the question of why similar strategies were not tried elsewhere, in the nonindustrializing cases. More importantly, it highlights the role of (and space for) innovation in relaxing political constraints that might have otherwise appeared irremovable.

Dual-Track Reform in China

During the 1970s, China was a centrally planned economy in which administered prices were a mechanism of generating rents and transfers to groups favored by the Communist regime. Price liberalization and the removal of obligatory grain deliveries to the state would generate significant efficiency gains in the countryside, where the bulk of the population lived. But it would come at the expense of depriving the state of its tax base, and urban workers of their cheap rations of food. By the standards of basic political economy models, these strong redistributive consequences provide an adequate explanation of why efficiency-enhancing reforms were resisted by the Chinese leadership.

But the Chinese government was able to devise a shortcut. Starting in the late 1970s, it made use of policy innovations such as two-track pricing and special economic zones that effectively delinked market-oriented incentives from their usual distributive implications. Consider, for example, how agriculture was reformed. Instead of abolishing the planned grain deliveries at fixed prices, the state simply grafted a market system on top of the centralized allocation system. Once the planned deliveries were made at state-set prices, farmers were free to sell additional amounts at any price the market would bear. As Lau, Qian, and Roland (2000) show, this system delivers allocative efficiency under fairly nonrestrictive conditions. But from a political economy perspective, the main virtue of the dual-track approach was that it shielded the prevailing stream of rents from the effects of the reform. The state did not lose its revenue, and urban workers were not denied their cheap food rations.

China's special economic zones functioned similarly. Rather than liberalize its trade regime in the standard way, which would have decimated the country's inefficient state enterprises, China allowed firms in special economic zones to operate under near-free-trade rules while maintaining trade restrictions elsewhere until the late 1990s. This enabled China to insert itself in the world economy while protecting employment and rents in the state sector. The Chinese Communist Party was strengthened and enriched, rather than weakened, as a result.

Democratization in South Africa

The black majority demanding democracy from the minority apartheid regime in South Africa faced a classic political economy problem. Both sides understood that once the African National Congress (ANC) obtained power, it would come under strong pressure from the black majority it represented to expropriate (or at least severely tax) the white elite. For the latter to accede to political reform, it had to have credible guarantees against expropriation. In view of the international sanctions and the economic decline they faced, the elites would have been better off under democracy—but only provided that moderate future taxation could be assured. In the absence of such guarantees, it remained in the elites' interest to keep suppressing the black majority even at substantial economic cost to themselves and the country.

Nelson Mandela was keenly aware of the problem: "Especially in the first few years of the democratic government," he said in 1991 (as quoted in Inman and Rubinfeld 2012, p. 784), "we may have to do something to show that the system has got an inbuilt mechanism which makes it impossible for one group to suppress the other." In the run-up to the democratic transition of 1994, South Africa's federal institutions were specifically designed to prevent the expropriation of the rich white minority by the poor black majority. Two key provisions were critical. First, critical redistributive services were left in the hands of provincial authorities.

Second, borders ensured at least one important province (Western Cape) would remain in the hands of the white minority. Inman and Rubinfeld (2012) argue these two arrangements together created a "hostage game" in which the incentives of a black national government to tax the white elites were moderated by the implicit threat of the local authorities in the Western Cape to respond by reducing service provision to the blacks in their province. Creative manipulation of the rules enabled both a political transition and a movement closer to the efficiency frontier.

Other Examples

One can greatly multiply these examples. During the 1980s and 1990s, reformist technocrats in Latin America overcame opposition from powerful insider groups by packaging liberalization and privatization (with strong redistributive effects) along with disinflation programs perceived by most—elites included—as inevitable and necessary (Rodrik 1994). In the United States, Trade Adjustment Assistance and other measures that operate as social insurance and compensation are the usual sweeteners offered to labor groups to buy their support for international trade agreements (Destler 2005). The US Congress relented on allowing the auction of radio frequencies only when political strings were devised—limiting the auctions to commercial wireless services and granting special rights to specific groups (women, minorities, small businesses)—to ensure that Congressional members would derive specific advantages from the move. Leighton and López (2013, p. 147) write that "in the end, everyone with a decision-making role in Congress got something, either more revenues or more political oversight."

These strategies represent innovations that shift the political transformation frontier out. They enable the capture of efficiency gains in ways that conserve the power of insiders and elites and protect their rents. Sometimes they enable radical political change, as in the South African case. At other times, they are designed to preclude political change, as in China. And even though I have focused on large-scale policy innovations that changed the course of nations, one can easily come up with a long list of others that are less revolutionary: the income tax, old-age pensions, the most-favored-nation principle in international trade, bank deposit insurance, work requirements for welfare recipients, conditional cash transfers, central bank independence, marketable pollution trading. What these all have in common is that they unblock resistance to change to allow society to move closer to the efficient frontier.

Where Do Policy Ideas Come From?

What determines the development and use of innovative political strategies? Why are some political systems blessed with a greater abundance of political innovations? What explains the timing of their emergence? Just as in the case of technological innovation, we might not be able to provide full answers to such questions. Innovation occurs, in large part, as a result of serendipity, as fundamental scientific discoveries yield unanticipated practical benefits or as experimentation and trial and error result in new products and processes. Similarly, we must presume there is a strong idiosyncratic element in political leadership and political creativity. Nevertheless, as the economic literature on research and development and endogenous growth indicates, certain systematic elements are also in play (Segerstrom, Anant, and Dinopoulos 1990; Aghion and Howitt 1998). For example, technological innovation responds to market incentives—the pursuit of monopoly profits through the acquisition of temporary advantages over competitors. Likewise, policy ideas that relax political constraints can be thought of as the consequence of both idiosyncratic processes and purposive behavior. Taking political ideas seriously opens up an interesting new line of research. Here I sketch, in broad outlines, some sources of new ideas that may be amenable to systematic analysis.

Political Entrepreneurship

Inefficiency creates opportunities for political entrepreneurship. As long as there are unexploited efficiency gains to be had, political agents have *some* incentive to engage in such search, regardless of the specific motives animating them. Economists, for example, develop proposals that they think will enhance economic performance. Sometimes (although not always, as Acemoglu and Robinson 2013 emphasize), these proposals take political feasibility into account. But ultimately, political entrepreneurs are the ones who arbitrage between academic ideas and political inefficiencies. It would be nice to know the circumstances under which such arbitrage actually takes place and political entrepreneurs are actually able to implement their policy innovations; for now, there seems to be little research addressing this question.

In their recent book, Leighton and López (2013) place special emphasis on political entrepreneurship in making policy reform possible. For new ideas to overcome vested interests, they write (p. 134), it must be the case that "entrepreneurs notice and exploit those loose spots in the structure of ideas, institutions, and incentives." They provide four case studies of this process: spectrum license auctions, airline deregulation, welfare reform, and housing finance. In their words (p. 178): "[T]he public face of political change may be that of a madman, an intellectual, or an academic scribbler. But whatever form these leaders may take, they are political entrepreneurs—people whose ideas and actions are focused on producing change." As these authors stress, political entrepreneurship can be socially harmful, as when the pursuit of individual rents comes at the expense of overall inefficiency. But the returns from shifting the political transformation frontier out can be very large as well.

Learning-by-doing

Entrepreneurship is linked to learning. Just as firms travel down their cost curves as a result of accumulated experience, public organizations such as bureaucracies can learn about opportunities to reap efficiency gains. A large literature examines the potential trade-off between learning and obsolescence as organizations age (Sørensen and Stuart 2000). Similarly, politicians might learn from their past successes and failures. The evolutionary approach to economics, based on trial and error by boundedly rational agents, provides a useful complementary perspective on learning, which also remains unexploited in political economy (Nelson and Winter 2002).

Technological learning often spills over to other firms, depressing the incentives for technological innovation. An interesting possibility is that political learning-by-doing is characterized by a similar externality. Political incumbents may be deterred from experimentation because they will bear the full cost of failed policy experiments, but will share the rents resulting from any successes with potential challengers who act as copycats. In this framework, more contestable political systems, allowing freer entry, may have ambiguous effects on political ideas. More competition means more entrepreneurs vying for new ideas. But it also means more copycats—political opponents waiting in the wings—reducing the incentive for experimentation and learning about the strategies that relax political constraints.

Policy Mutations

By "policy mutation," I refer to unplanned policy experimentation that arises along the margins of existing policies. Such experimentation often results from the inability of policymakers to implement prevailing rules to the letter, for administrative or other reasons. As with random mutations, these variations on generally accepted practice can generate new and improved policies by demonstrating better practical results. For example, the idea for dual-track policies in China arose not from the planners themselves, but from black markets in the Chinese countryside where farmers sold grain illegally. Planners were simply wise enough to understand that these markets-at-the-margin enriched farmers without harming the state, as long as the plan quotas themselves were enforced, and then to build public policy on that understanding. Similarly, experiments with "supersaver fares" in California and Texas greatly facilitated US airline deregulation during the 1970s by revealing the sizable price benefits of greater competition and freer entry (Leighton and López 2013, pp. 155–56).

Leitzel (2003) has written insightfully on the reformist consequences of what he calls "rule evasion." As he notes, "evasive behavior in essence presents an experiment, an alternative way of arranging society" (p. 23). Leitzel discusses two reasons why rule evasion paves the ground for new policies. First, the evasion typically becomes common knowledge and conveys a sense that the existing policy is a failure. Second, it creates incentives for reform either by suggesting an alternative to current policy (say, legalizing black markets) or creating a constituency for the reform. In terms of the argument in this paper, it is a source of ideas for policymakers about what can work better within political constraints.

Crises

Times of crises are occasions for reconsidering existing policies. This is both because prevailing interests may lose some of their legitimacy and because incumbents may be open to trying new remedies. The need for a new narrative is greater and so is the willingness to experiment. "In moments of uncertainty," writes Blyth

(2007, p. 762), "crisis-defining ideas not only tell agents 'what has gone wrong' but also 'what is to be done'."

In the United States, the Great Depression proved a veritable laboratory for novel institutional arrangements. Franklin D. Roosevelt (1932) uttered a famous call for "bold, persistent experimentation": "It is common sense to take a method and try it: If it fails, admit it frankly and try another. But above all, try something." To a much lesser extent, the inflationary crisis of the 1970s played a similar role, preparing the groundwork for new ideas in macroeconomics such as rational expectations and central bank independence. The recent financial crisis has made taxation and control of international capital flows more palatable, although the extent to which financial interests have been weakened remains debated.

While the association between crises and new ideas seems plausible, much remains to be explained. Why are some crises much more prone to new ideas? What explains the type of ideas that take hold? The Great Depression spawned the New Deal in the United States, fascism in some parts of Europe, and socialism in some other parts of Europe. Were these outcomes preordained by the structure of interests? To what extent did political entrepreneurship and ideas play an autonomous role?

Emulation

Perhaps the single most important source of ideas and policy innovation are practices that prevail elsewhere. The fact that a policy has worked—or at least is perceived to have worked—somewhere can be a powerful reason to copy it. Social security privatization in Chile, microfinance in Bangladesh, conditional cash grants in Mexico, and special economic zones in China are some examples of policy innovation that gained adherents around the world following implementation in their native settings. Much legal and regulatory reform in the developing world is modeled after existing models in North America or Western Europe. The appeal of "imported ideas" is clear. Ready-made policies eliminate or reduce the cost of home-grown innovation and experimentation. The perception of their success elsewhere can also act as a counterweight to powerful vested interests at home.

Of course, there is no guarantee that policy emulation will result in success. Context matters. Imported ideas can backfire because of ill fit with either the local economic or political landscape. Furthermore, emulation can be driven by bad motives as well as good ones. It may be used to provide aid donors with cover in case of failure, as a signal for new governments that they are the "good guys," and by domestic lobbies to legitimize their own self-interested agenda (Weyland 2008). DiMaggio and Powell (1983) have coined the term "isomorphic mimicry" to denote the pressures that organizations face to become similar, even as they struggle to change. Andrews (2013) documents how reform in poor countries through "isomorphic mimicry" results in the semblance of change, with little real progress achieved: a bureaucracy gets reorganized to look like those from advanced countries, but bureaucratic efficiency hardly improves. In Mukand and Rodrik (2005), my coauthor and I develop a formal model of the incentives for governments to mimic other countries' policies: implementing policies with a poor fit is costly, but so is experimentation—and an imperfectly informed electorate may be more likely to interpret domestic experimentation as an attempt at rent seeking, while being willing to accept emulation.

Conclusion: What Do Economists and Policy Analysts Gain by Considering the Role of Ideas?

The main argument of this paper is that, for all the emphasis placed on them in contemporary models of political economy, vested interests play a considerably less-determining role than appears at first sight. Indeed, because of their neglect of ideas, political economy models often do a poor job of accounting for policy change. There is frequently an after-the-fact feel to this brand of theorizing: if reform happens despite vested interests, it must be because those interests were not sufficiently well entrenched to begin with or reform didn't hurt them. Conventional models of policy stasis are incomplete if they sidestep the ideas that political agents have about strategies they can pursue. And they cannot fully shed light on reform when it does occur.

It is instructive to contrast my argument with that of Acemoglu and Robinson (2013), who argue in this journal that well-meaning reforms often fail or produce unintended consequences because they overlook the changes in political equilibrium the reforms generate. In much of policy advice, they write, politics is "largely absent from the scene." Acemoglu and Robinson maintain that "economic analysis needs to identify, theoretically and empirically, conditions under which politics and economics run into conflict, and then evaluate policy proposals taking this conflict and the potential backlashes it creates into account." I agree with them on the need to take politics into account. But Acemoglu and Robinson take vested interests largely as given, and as a result, they are rather pessimistic about what policy can achieve. In contrast, I have argued that successful policy ideas work precisely because they take politics into account. I suggest that it is possible to do better than simply avoid political conflicts; ideas can be useful to *relax* political constraints. Just as ill-conceived economic ideas can produce disastrous political effects, politically well-informed ideas can move us closer to the efficiency frontier in a manner that is consistent with underlying political realities.

Taking ideas seriously renders the notion of interests slippery and ephemeral. From the conventional political-economy standpoint, it is puzzling to observe instances in which elites resist reforms strenuously until the change actually happens, and then benefit from the reforms. The Korean military dictator President Park Chung-Hee threw the country's leading businessmen in jail when he came to power in 1961; they were released only after Park extracted promises from them that they would each undertake specific industrial investments. Given how the Korean economy prospered, these businessmen were hardly worse off for those investments. Similarly, the Chinese Communist leadership was among the main beneficiaries of the dual-pricing regime and other market-oriented policy innovations that it had refused to consider until Mao's death. The critical change in these instances was not a transformation in the structure of power, but the implementation of new ideas by those in power. Indeed, reform often happens not when vested interests are defeated, but when different strategies are used to pursue those interests, or when interests themselves are redefined.

Raising the profile of ideas in political economy models would also help alleviate the tension that exists today between political economy, on the one hand, and normative economics and policy analysis, on the other. Political economy seeks to explain political-economic outcomes. However, if policy outcomes are pinned down by the structure of interests, it is futile to make policy recommendations: there will be no takers for the recommendations, and such recommendations will be of no consequence. At best, they will constitute ideological fodder for vested interests, used to sweeten their exercise of naked power before the general public. When political economy becomes too enamored of vested interests at the expense of ideas, social science squeezes normative policy analysis out of useful existence. An explicit consideration of the role of ideas would free up some space for policy analysis.

Finally, a focus on ideas provides us with a new perspective on vested interests too. As social constructivists like to put it, "interests are an idea." Even if economic actors are driven purely by interests, they often have only a limited and preconceived idea of where their interests lie. This may be true in general, of course, but it is especially true in politics, where preferences are tightly linked to people's sense of identity and new strategies can always be invented. What the economist typically treats as immutable self-interest is too often an artifact of ideas about who we are, how the world works, and what actions are available.

■ *I thank seminar participants at Harvard's Kennedy School and the Elinor Ostrom workshop at Indiana University for comments, David Autor, Chang-Tai Hsieh, Avinash Dixit, and Timothy Taylor for detailed suggestions, Sharun Mukand for helpful conversations along the way, and Mark Blyth for reading recommendations. I owe a special debt to the recent book by Edward López and Wayne Leighton (2012) for stimulating me to put down on paper a number of ideas I had been mulling over for some time.*

References

Acemoglu, Daron. 2003. "Why Not a Political Coase Theorem? Social Conflict, Commitment, and Politics." *Journal of Comparative Economics* 31(4): 620–52.

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz. 2009. "Fragility of Asymptotic Agreement under Bayesian Learning." Unpublished paper, February. http://economics.mit.edu/files/3795.

Acemoglu, Daron, and James A. Robinson. 2006.

"Economic Backwardness in Political Perspective." *American Political Science Review* 100(1): 115–31.

Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York: Crown.

Acemoglu, Daron, and James A. Robinson. 2013. "Economics versus Politics: Pitfalls of Policy Advice." *Journal of Economic Perspectives* 27(2): 173–92.

Aghion, Philippe, and Peter Howitt. 1998. *Endogenous Growth Theory.* Cambridge, MA: MIT Press.

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics* 115(3): 715–53.

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives* 19(1): 9–32.

Alesina, Alberto, Guido Cozzi, and Noemi Mantovan. 2012. "The Evolution of Ideology, Fairness and Redistribution." *Economic Journal* 122(565): 1244–61.

Alesina, Alberto, and Howard Rosenthal. 1995. *Partisan Politics, Divided Government, and the Economy.* Cambridge University Press.

Andrews, Matt. 2013. *The Limits of Institutional Reform in Development: Changing Rules for Realistic Solutions.* Cambridge University Press.

Atran, Scott, and Jeremy Ginges. 2012. "Religious and Sacred Imperatives in Human Conflict." *Science* 336(6083): 855–57.

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

Béland, Daniel, and Robert Henry Cox, eds. 2011. *Ideas and Politics in Social Science Research.* Oxford University Press.

Blyth, Mark. 2002. "Institutions and Ideas." In *Theory and Methods in Political Science,* edited by Dave Marsh and Gerry Stoker, pp. 292–311. London: Macmillan.

Blyth, Mark. 2007. "Powering, Puzzling, or Persuading? The Mechanisms of Building Institutional Orders." *International Studies Quarterly* 51(4): 761–77.

Blyth, Mark. 2011. "Ideas, Uncertainty, and Evolution." Chap. 4 in *Ideas and Politics in Social Science Research,* edited by Daniel Béland and Robert Henry Cox. Oxford University Press.

DellaVigna, Stefano, and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics* 122(3): 1187–1234.

Denzau, Arthur T., and Douglass C. North. 1994. "Shared Mental Models: Ideologies and Institutions." *Kyklos* 47(1): 3–31.

Destler, I. M. 2005. *American Trade Politics.* 4th ed. Washington, DC: Peterson Institute for International Economics.

DiMaggio, Paul J., and Walter W. Powell. 1983. "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields." *American Sociological Review* 48(2): 147–60.

Dixit, Avinash K. 1996. *The Making of Economic Policy: A Transaction-Cost Politics Perspective.* Cambridge, MA: MIT Press.

Dixit, Avinash K., and Jörgen Weibull. 2007. "Political Polarization." *PNAS* 104(18): 7351–56.

Elster, Jon. 2000. "Rational Choice History: A Case of Excessive Ambition." *American Political Science Review* 94(3): 685–95.

Fernandez, Raquel, and Dani Rodrik. 1991. "Resistance to Reform: Status Quo Bias in the Presence of Individual-Specific Uncertainty." *American Economic Review* 81(5): 1145–55.

Goldsmith, Jack L., and Eric A. Posner. 2005. *The Limits of International Law.* Oxford and New York: Oxford University Press.

Grossman, Gene M., and Elhanan Helpman. 1994. "Protection for Sale." *American Economic Review* 84(4): 833–50.

Hammond, Peter J. 1979. "Straightforward Individual Incentive Compatibility in Large Economies." *Review of Economic Studies* 46(2): 263–82.

Hay, Colin. 2011. "Ideas and the Construction of Interests." Chap. 4 in *Ideas and Politics in Social Science Research,* edited by Daniel Béland and Robert Henry Cox. Oxford University Press.

Howse, Robert. 2013. "Thucydides and Just War: How to Begin to Read Walzer's *Just and Unjust Wars.*" *European Journal of International Law* 24(1): 17–24

Inman, Robert P., and Daniel L. Rubinfeld. 2012. "Federal Institutions and the Democratic Transition: Learning from South Africa." *Journal of Law, Economics & Organization* 28(4): 783–817.

Johnson, Simon, and James Kwak. 2010. *13 Bankers: The Wall Street Takeover and the Next Financial Meltdown.* New York: Pantheon Books.

Krueger, Anne O. 1974. "The Political Economy of the Rent-Seeking Society." *American Economic Review* 64(3): 291–303.

Krueger, Anne O. 1997. "Trade Policy and Economic Development: How We Learn." *American Economic Review* 87(1): 1–22.

Laffont, Jean-Jacques, and Jean Tirole. 1991. "The Politics of Government Decision Making: A Theory of Regulatory Capture." *Quarterly Journal of Economics* 106(4): 1089–1127.

Lau, Lawrence, J., Yingyi Qian, and Gérard Roland. 2000. "Reform without Losers: An Interpretation of China's Dual-Track Approach to Transition." *Journal of Political Economy* 108(1): 120–43.

Leighton, Wayne, and Edward López. 2013. *Madmen, Intellectuals, and Academic Scribblers: The*

Economic Engine of Political Change. Stanford, CA: Stanford University Press.

Leitzel, Jim. 2003. *The Political Economy of Rule Evasion and Policy Reform*. London and New York: Routledge.

Mehta, Pratap Bhanu, and Michael Walton. 2012. "India's Political Settlement and Development Path." Unpublished paper, December.

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.

Nelson, Richard R., and Sidney G. Winter. 2002. "Evolutionary Theorizing in Economics." *Journal of Economic Perspectives* 16(2): 23–46.

Olson, Mancur. 1993. "Dictatorship, Democracy, and Development." *American Political Science Review* 87(3): 567–76.

Pande, Rohini. 2011. "Can Informed Voters Enforce Better Governance? Experiments in Low-Income Democracies." *Annual Review of Economics*, vol. 3, pp. 215–37.

Peltzman, Sam. 1976. "Toward a More General Theory of Regulation." *Journal of Law and Economics* 19(2): 211–40.

Rodrik, Dani. 1994. "The Rush to Free Trade in the Developing World: Why So Late? Why Now? Will It Last?" Chap. 3 in *Voting for Reform: Democracy, Political Liberalization, and Economic Adjustment*, edited by Stephen Haggard and Steven B. Webb. New York, NY: Oxford University Press for the World Bank.

Rodrik, Dani. 1995. "The Political Economy of Trade Policy." Chap. 28 in *Handbook of International Economics*, vol. 3, edited by Gene M. Grossman and Kenneth Rogoff. Amsterdam: North-Holland.

Roosevelt, Franklin D. 1932. Oglethorpe University Commencement Address, May 22. http://newdeal.feri.org/speeches/1932d.htm.

Ruggie, John Gerard. 1998. "What Makes the World Hang Together? Neo-Utilitarianism and the Social Constructivist Challenge." *International Organization* 52(4): 855–85.

Schultz, Theodore W. 1964. *Transforming Traditional Agriculture*. New Haven: Yale University Press.

Segerstrom, Paul S., T. C. A. Anant, and Elias Dinopoulos. 1990. "A Schumpeterian Model of the Product Life." *American Economic Review* 80(5): 1077–91.

Sen, Amartya. 2006. *Identity and Violence: The Illusion of Destiny*. New York: W. W. Norton.

Sørensen, Jesper B., and Toby E. Stuart. 2000. "Aging, Obsolescence, and Organizational Innovation." *Administrative Science Quarterly* 45(1): 81–112.

Stigler, George J. 1971. "The Theory of Economic Regulation." *Bell Journal of Economics and Management Science* 2(1): 3–21.

Wendt, Alexander. 1999. *Social Theory of International Politics*. Cambridge University Press.

Weyland, Kurt. 2008. "Toward a New Theory of Institutional Change." *World Politics* 60(2): 281–314.

Yanagizawa-Drott, David. 2012. "Propaganda and Conflict: Theory and Evidence from the Rwandan Genocide." CID Working Paper 257, Center for International Development, Harvard University.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 209–234

An Economist's Guide to Visualizing Data

Jonathan A. Schwabish

Once upon a time, a picture was worth a thousand words. But with online news, blogs, and social media, a good picture can now be worth so much more. Economists who want to disseminate their research, both inside and outside the seminar room, should invest some time in thinking about how to construct compelling and effective graphics.

An effective graph should tap into the brain's "pre-attentive visual processing" (Few 2004; Healey and Enns 2012). Because our eyes detect a limited set of visual characteristics, such as shape or contrast, we easily combine those characteristics and unconsciously perceive them as an image. In contrast to "attentive processing"—the conscious part of perception that allows us to perceive things serially—pre-attentive processing is done in parallel and is much faster. Pre-attentive processing allows the reader to perceive multiple basic visual elements simultaneously. Here is a simple example; count the occurrences of the number 3 in the following set:

1269548523612356987458245
0124036985702069568312781
2439862012478136982173256

Now repeat the task with this set of numbers:

126954852**3**612**3**56987458245
01240**3**6985702069568**3**12781
24**3**98620124781**3**698217**3**256

■ *Jonathan A. Schwabish is a Principal Analyst, Congressional Budget Office, Washington, DC. His email address is jonathan.schwabish@cbo.gov.*

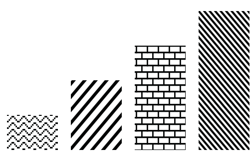
http://dx.doi.org/10.1257/jep.28.1.209 doi=10.1257/jep.28.1.209

The instances in the second set are much easier to find because they are encoded using a different pre-atttentive attribute—in this case, the intensity of boldface type.

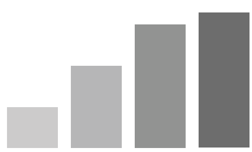
It takes imagination to create high-quality images that illustrate data accurately and effectively and also show some understanding and appreciation of how people acquire information. Indeed, what is known as the "Picture Superiority Effect" refers to our ability to retain more information seen through pictures than through words (for example, Medina 2008; Hockley and Bancroft 2011). There are thousands of approaches to presenting data: for starters, consider the vast information on how to choose fonts, colors, styles, layouts, and chart types. Three basic principles seem especially useful.

First, *show the data*. People read graphs in a research report, article, or blog to understand the story being told. The data are the most important part of the graph and should be presented in the clearest way possible. But that does not mean that *all* of the data must be shown—indeed, many graphs show too much.

Second, *reduce the clutter*. Chart clutter, the use of unnecessary or distracting visual elements, will tend to reduce effectiveness. Clutter comes in dark or heavy gridlines; unnecessary tick marks, labels, or text; unnecessary icons or pictures; ornamental shading and gradients; and unnecessary dimensions. Too often graphs use textured or filled gradients



when simple shades of a color could accomplish the same task.



In some cases, familiar data markers (■◆●×) are used to distinguish between several data series on a graph, but when the markers intersect and overlap they end up cluttering the patterns.

Third, *integrate the text and the graph*. Standard research reports often suffer from the "slideshow effect," in which the writer narrates the text elements that appear in the graph. A better model is one in which visualizations are constructed to complement the text and at the same time to contain enough information to stand alone (Corum 2013). As a simple example, legends that define or explain a line, bar, or point are often placed far from the content of the graph—off to the right or below the graph. Integrated legends—right below the title, directly on the chart, or at the end of a line—are more accessible.

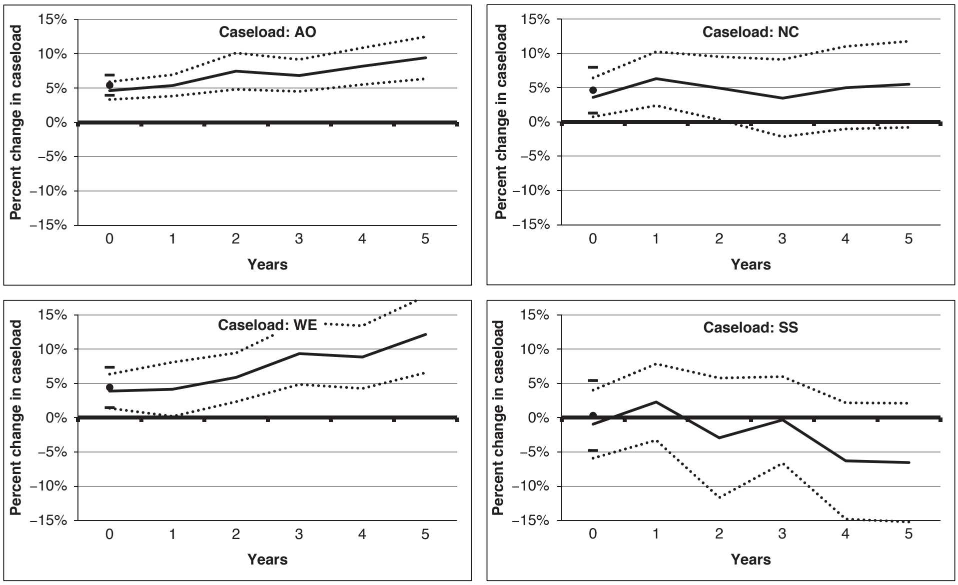
These three principles embody the idea that an author should support the reader's acquisition of information quickly and easily. By stripping out unnecessary clutter, emphasizing the data, and using certain pre-attentive attributes (for example, hue (color), size, orientation, and shape) graphs can more clearly and more effectively communicate information. However, default graph options in many statistical programs tend to add clutter and to separate text and graphs, and so researchers need to consider overriding those defaults and perhaps adding annotation to create graphs that communicate information more effectively to the reader.

This article encourages economists to think more strategically about how to visualize their data and presents some pathways to create better, more effective graphs. The next section demonstrates the principles in eight graphs remade using nothing more complicated than Excel. The discussion then addresses some types and purposes of different data visualizations, and briefly reviews some tools and sources of information that researchers can use to improve their existing graphs or to create new ones. One thing researchers should keep in mind is that graphs in research reports or articles, and even those shown in verbal presentations, are not meant for the author, but for the reader or the seminar audience. The line chart that a researcher uses in the data exploration phase—with default gridlines, tick marks, and colors—may not be the one that will best communicate the researcher's ideas to others. Discussions of data visualization are only now making their way into economics journals and conferences, but perhaps this is just the beginning and our discipline's understanding of the importance of good visualization will expand and grow.

Eight Graphs Transformed

Poor graphs communicate ineffectively, or even worse, provide a distorted impression of the data. This section shows how eight graphs could be redesigned to demonstrate the application of the three principles outlined above. Some decisions are subjective, of course—line thickness, series order, axis label style. Other decisions, I would argue, are just objectively *better* ways to convey meaning. All of the redesigned graphs were constructed in Excel and required slight variations from the program's default settings. Garamond—a classic serif font—is used to slightly distinguish the graphs from the *Journal of Economic Perspectives*'s Baskerville typeface. The electronic version of the *JEP* often uses color, which can be an important tool in data visualization to invoke emotion, emphasize graphical elements, or simply add aesthetic value. The print *JEP* does not use color, but all graphs that use color in the electronic version of the *JEP* are designed to work in greyscale for print readers. The choice of a color palette, like the choice of a font, also can be subjective, but following some basic guidelines can improve communication. In the final section, I discuss some tools that can help with those strategies and selections.

Figure 1A
An Original Line Chart



Source: Klerman and Danielson (2011).

The Line Chart

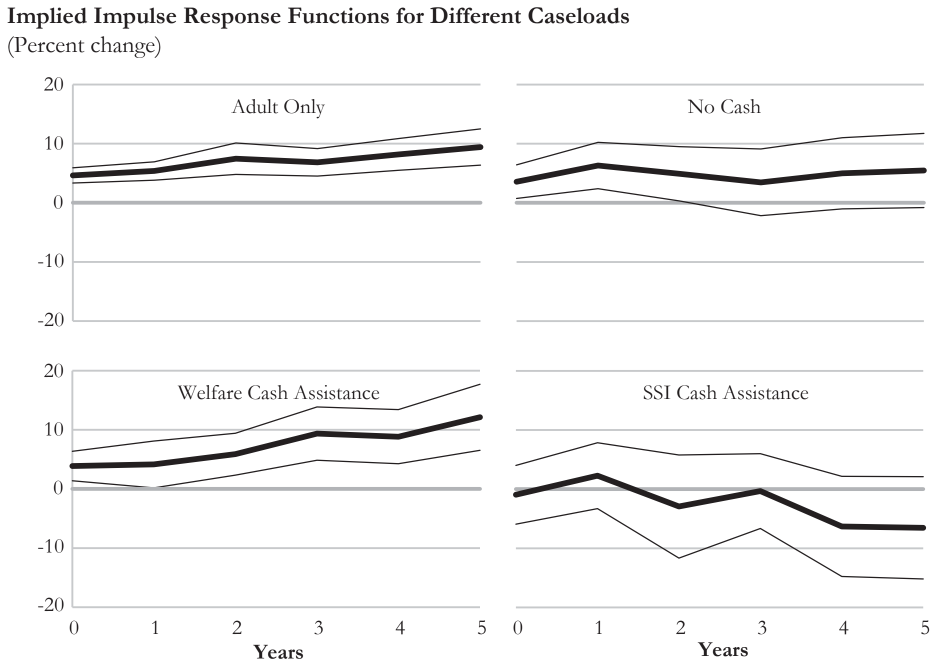
Figure 1A summarizes regression results of the correlation between the long-run unemployment rate in the United States and Supplemental Nutrition Assistance Program caseloads for four groups.[1] Instead of a single line chart with multiple series or four large, separate charts, the authors have used a "small multiples" approach in which four smaller charts are grouped together. Instead of packing as much information as possible into a single graph, researchers should probably use this approach more often. However, in other ways the graphs violate the three principles outlined above.

Perhaps most importantly, a graph should emphasize the data, but the darkest and thickest line on these graphs is the 0 percent gridline. Your eye is immediately drawn to that thick, horizontal gridline rather than to the important parts of the graph, namely the coefficient line and the standard errors. Also notice that the data values in the last point of the WE and SS charts actually exceed the 15 percent data marker—the graphs fail to show all data points.

Some elements add unneeded clutter. For example, the y-axis labels and percentage signs are redundant and add clutter (there are 28 percentage signs in all!) and the tick marks on the y-axes also seem unnecessary.

[1] The authors were kind enough to send me the data behind their figure for reproduction here.

Figure 1B
A Revised Line Chart



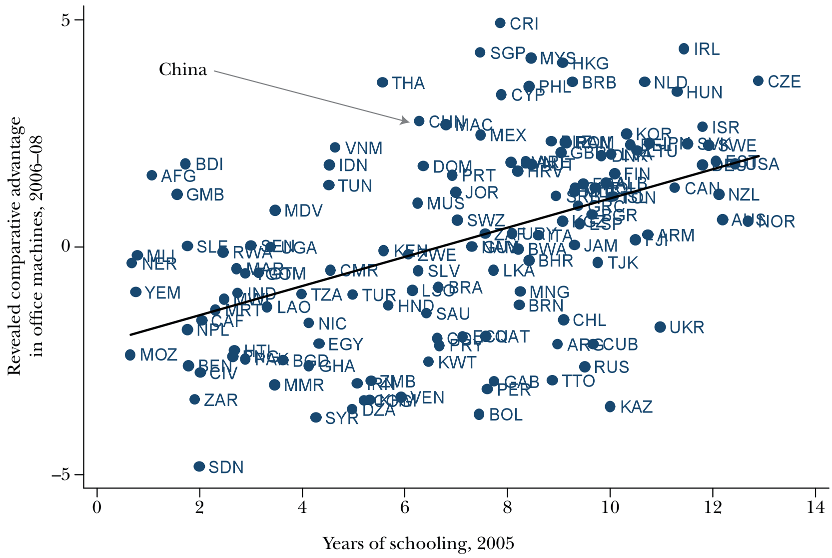
Finally, what do AO, NC, WE, and SS mean in the figure? In this article, these terms are explained on the third and fourth pages, 15 pages before this figure is presented. It seems unfair to ask a reader to search the paper to decode the meaning of those labels. Instead, the text and the graphs can be integrated.

Figure 1B offers an alternative version of this figure. The darkest line now shows the data (in this case, the coefficient estimate). The gridlines have been lightened, but leaving the 0 percent gridline slightly darker so it can provide a baseline for the series that dip below zero. Because the charts are aligned vertically and horizontally, I eliminated two sets of labels to reduce clutter and to help highlight the central message. I eliminated the percent signs and identified the unit in parentheses below the title. This last edit is essentially a style choice—there is nothing inherently wrong with using rotated text on the vertical axis, but it does require readers to turn the page sideways or tilt their heads (Robbins 2013a).

The title is now above the graph—some publishers place titles below graphs even though readers tend to start reading from the top left, move down along the left margin, and then move to the right. Repositioning the word "Caseload" into the title—used four times in the original—leaves room to spell out the abbreviations. One could, of course, argue about many of the choices illustrated here: Perhaps four sets of axis labels are better than two, or dotted standard error lines would be preferable to solid ones. But surely, a number of these changes are clear improvements on the original figure.

Figure 2A
A Clutterplot Example

Education and Exports of Office Machines



Source: Hanson (2012).

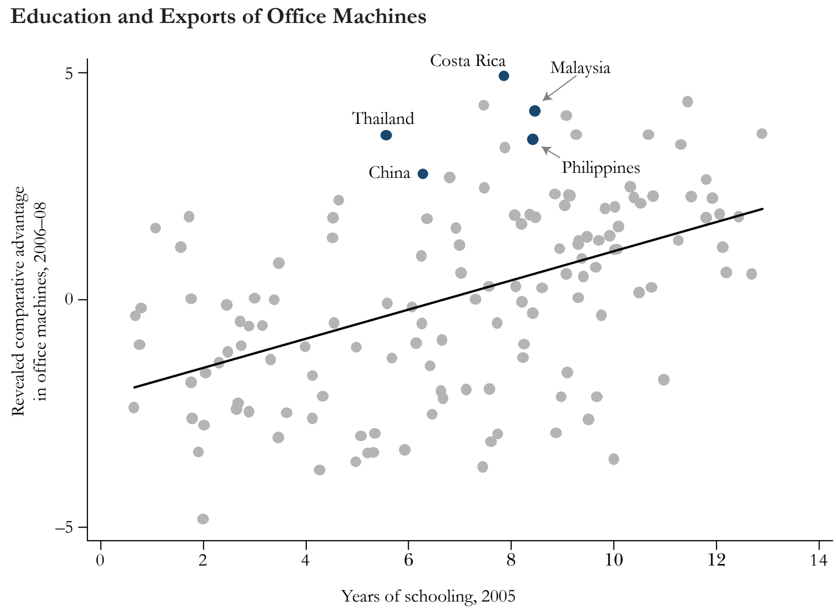
The Clutterplot

The next chart (Figures 2A and 2B) comes from the *JEP* (Hanson 2012, p. 54). The explanation in the original article is as follows (emphasis added):

> [The figure] plots countries' revealed comparative advantage in office machines . . . averaged over 2006 to 2008, against the average years of schooling of the adult population in 2005 . . . *China* is above the regression line, indicating that its specialization in the sector is greater than one would expect given its level of education, but it is hardly an extreme outlier. Other middle-income countries—including *Costa Rica*, the *Philippines, Malaysia*, and *Thailand*—have larger positive residuals.

Congratulations if you already know the three-letter codes for the five countries referenced in the text. Even if you do, try to find them in the haystack of labels and dots. The redesign offers an alternative that shows the data and reduces the clutter. I eliminated all labels other than those for the five countries under discussion, which I spelled out, and I made the five data points darker, thus deemphasizing the other points but still showing the important information.

Figure 2B
Revising the Clutterplot Example

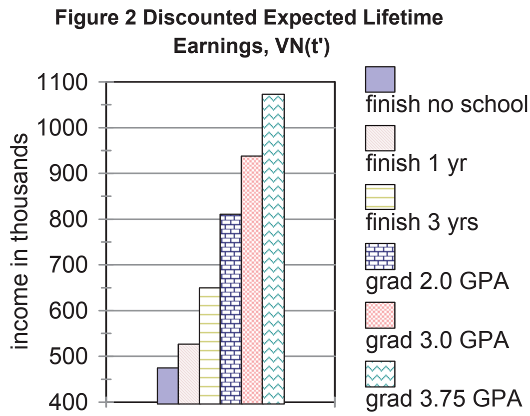


One potential objection to naming only some of the points is the concern that some readers may want to search through the figure for individual countries or identify various outlier points. There will often be some tension between presenting enough data to tell the story and presenting additional data that some readers might want. But most journals and many individual researchers now have their own websites, so posting the complete actual data for interested readers can be relatively simple and inexpensive. If it is important to include the data within the paper itself, a supplemental table in the text or as an appendix could be a more straightforward method of presentation.

The Basic Column Chart

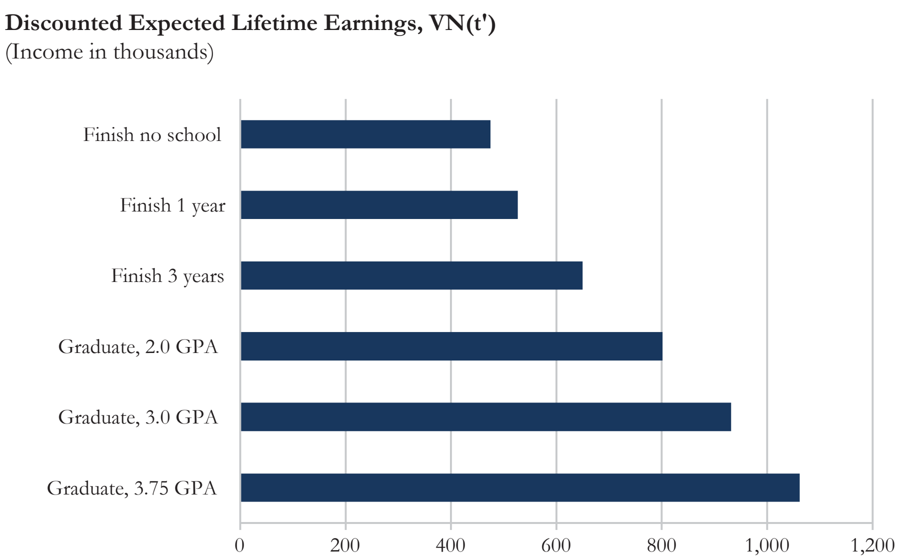
When it comes to bar and column charts, a first rule is to start the chart at zero. Otherwise, the differences between the columns are overemphasized. Figure 3A shows one such example of a column chart that does not start at zero. For example, if you look at the second-shortest bar, with a value just above 500, it certainly does not appear at a glance that it is half the length of the longest bar. This figure had different colors for each bar, which is not necessary but may be useful. The axis in the redesigned version, shown in Figure 3B, starts at zero and, to make room for the labels that are integrated with the chart, the figure is rotated horizontally. In this version, the second-shortest bar now appears at a glance to be about half the length of the longest bar.

Figure 3A
The Basic Column Chart



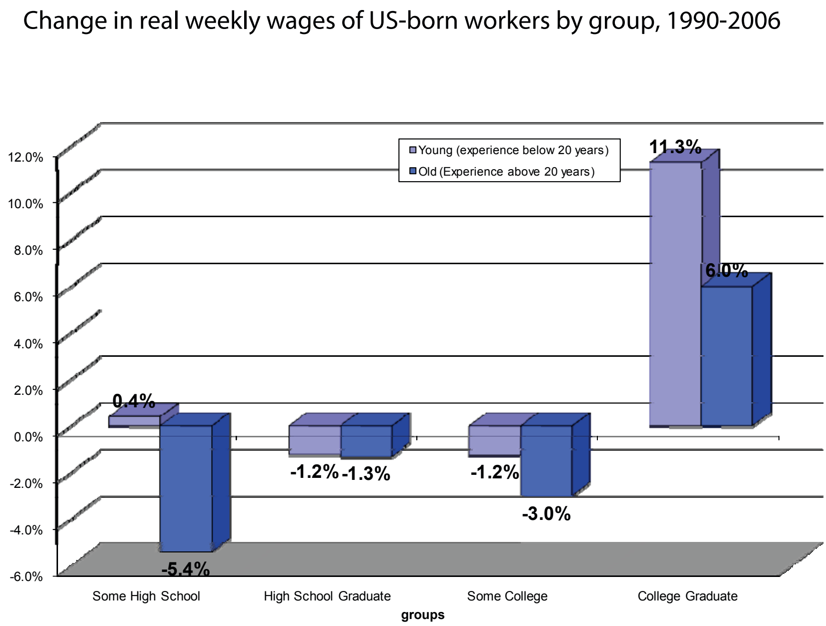
Source: Stinebrickner and Stinebrickner (2013).

Figure 3B
The Revised Column Chart



Source: Author's calculations using numbers inferred from text in Stinebrickner and Stinebrickner (2013).

Figure 4A
A 3D Chart



Source: Ottaviano and Peri (2008).

The 3D Chart

Figure 4A uses the now-familiar 3D effect. In such graphs, the third dimension does not plot data values, but it does add clutter to the chart and, worse, it can distort the information. Look at the far-right-hand bar, labeled 6 percent: No point of the column touches the gridline for that value. This software tool—like many others—uses perspective to give depth to the imaginary plane that runs across the top of the column, intersecting the gridline. But most readers will perceive the actual value of the column as less than 6 percent. Figure 4B shows a redesign: cancel the 3D treatment and integrate the disconnected legend with the graph. Notice that inserting the common baseline—portrayed in the original by a hovering, barely perceptible thin gray line—permits a more effective comparison among groups.

The Unbalanced Chart

The source material for Figure 5A originally appeared in an interactive visualization on the Organisation for Economic Co-operation and Development (OECD) website (http://www.oecd.org/gender/data/proportionofemployedwhoareseniormanagersbysex.htm); a static version was later reproduced in a *New York Times* Economix blog post (http://economix.blogs.nytimes.com/2013/04/02/comparing-the-worlds-glass-ceilings/?_r=2).

Figure 4B
Flattening a 3D Chart

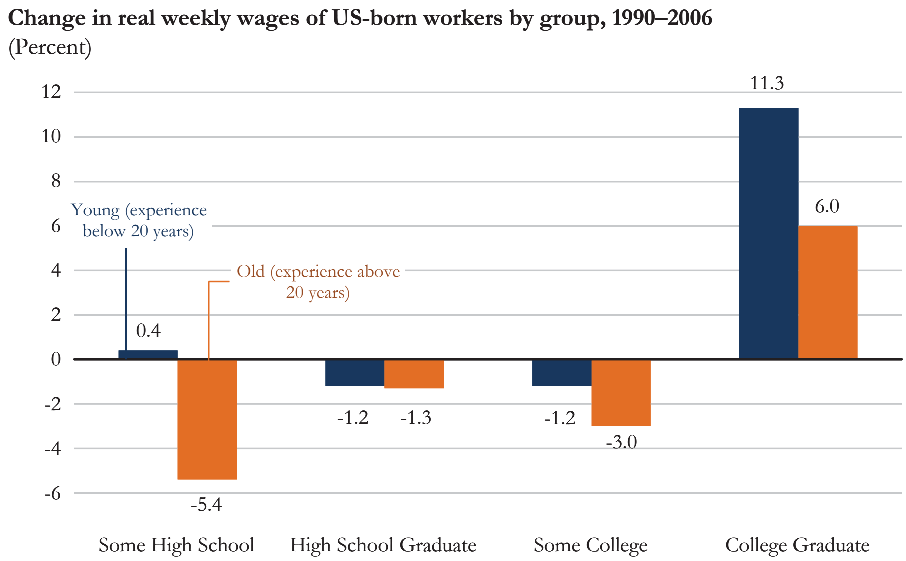
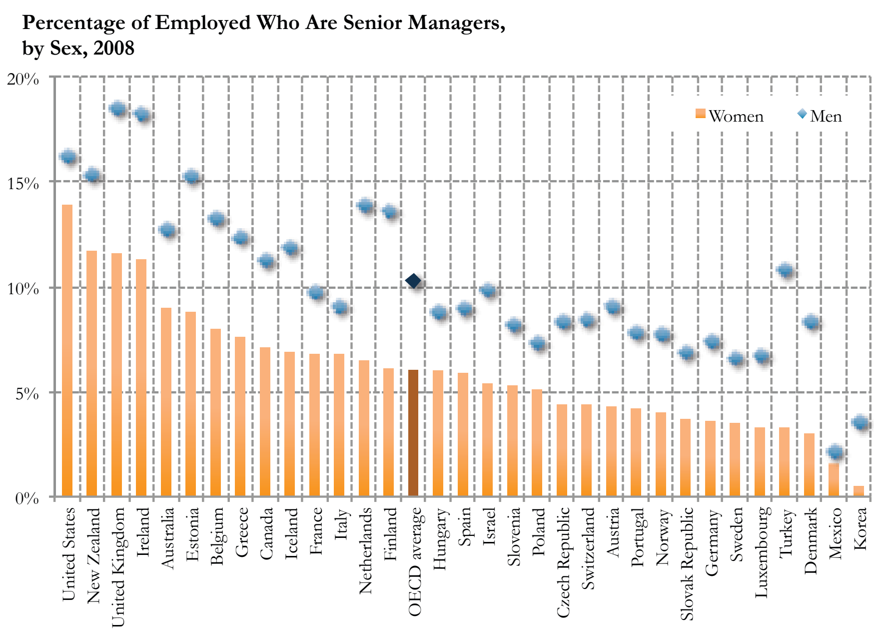


Figure 5A
An Unbalanced Chart



Source: Author, based on OECD (no date) and Rampell (2013).

The chart is an ineffective communication tool in either version for three main reasons. First, the same kinds of data are plotted using different types of encoding so that it is difficult to compare location (diamonds) with length (bars). Further, the relationship between the two values is obscured: The data points for men are too far away from those for women and there is no visual connection between them. The original versions used color—red (for women) and blue (for men) in the *New York Times* and orange and blue in the OECD version—which can serve to draw attention to or from certain elements.

Second, the columns for women take up a much larger proportion of the graph than do the diamonds for men, overemphasizing the data for women. If the intention is to give greater emphasis to the data for women, then a descriptive title—such as "Women's Employment as Senior Managers Averaged 6 Percent in 2008"—would have helped clarify the meaning. Furthermore, the gradient color shading in the columns is darker at the bottom than at the top where the data are truly encoded.

Finally, there are many gridlines, all heavy, and the percent signs on the y-axis labels are redundant. Additionally, the x-axis labels are potentially difficult to scan because they are vertical.

The chart could be redesigned in a number of ways. Paired bars could be used for men and women, or the figure could be turned into a table (Schwabish 2013a shows some other alternatives). Figure 5B shows a less-common type of visual approach. For some readers, these different types of plots may not be perceived as quickly as more commonly used graph types, such as bars or lines, but it's instructive to remember that scatterplots, not so long ago a novelty in mainstream publishing, now appear regularly. Just as our text literacy can expand with experience and exposure, so can our graphic literacy.

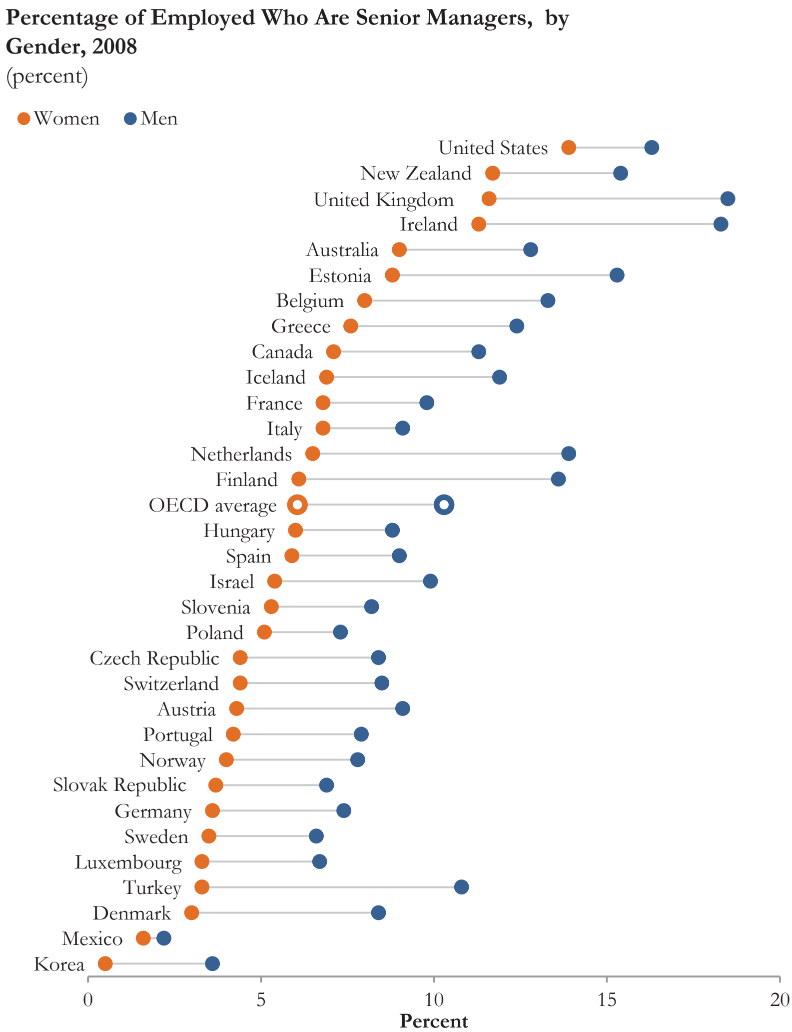
The redesigned chart shown in Figure 5B has several characteristics worth noting. First, the data are encoded similarly for men and women so that the reader can better perceive the connection between the two series and compare them. Second, the title, units, and legend are integrated and placed at the top-left of the chart, which helps the reader "enter" the chart. Third, the country labels are rotated horizontally and incorporated directly onto the chart with the thin gray connecting lines helping to illustrate the comparison between data points for men and women. Finally, the average value for the OECD as a whole is an unfilled circle (in a version where black and white printing is not considered, different shades or colors could be used instead).

One potential shortcoming of the redesign is the lack of vertical gridlines. In the original, the gridlines perhaps helped distinguish more specific values. There is often a tension in this aspect of data visualization—how much of the purpose of the graph is to explain an idea and how much is to provide specific data (Few 2005; Schwabish 2013b)?

The Spaghetti Chart

If a line chart has too many series, any single trend will be obscured. Such charts are sometimes called "spaghetti charts" (Nussbaumer 2013). If too much

Figure 5B
Revising the Unbalanced Chart

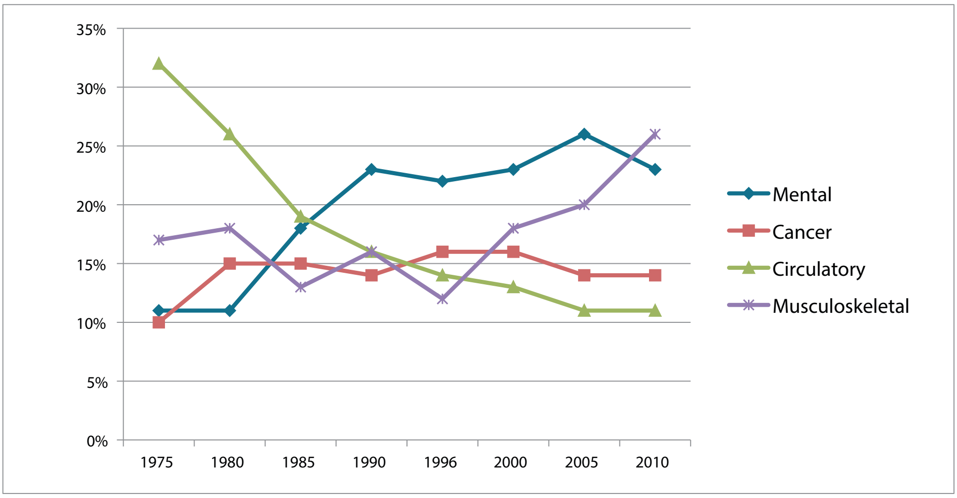


information is plotted, readers can have difficulty pulling out the meaning of a single series or drawing an overall conclusion. Figure 6A is not an extreme example of a spaghetti chart, with only four lines. Nonetheless, consider some of its shortcomings: First, the data markers on every point make it difficult to pre-attentively follow any single series. Second, the legend is located far from the data and the order of the legend does not match the order of the lines.

One alternative to the spaghetti chart is to create smaller charts in series (sometimes called sparklines or small multiples; see Tufte 2006). Instead of a single, dense presentation of data, Figure 6B splits the information into four separate, smaller

Figure 6A
A Spaghetti Chart

27. Initial DI Worker Awards by Major Cause of Disability—Calendar Years 1975-2010



Source: Social Security Advisory Board (2012).

Figure 6B
Revising the Spaghetti Chart

**Initial DI Worker Awards by Major Cause of Disability—
Calendar Years 1975–2010**
(Percent)

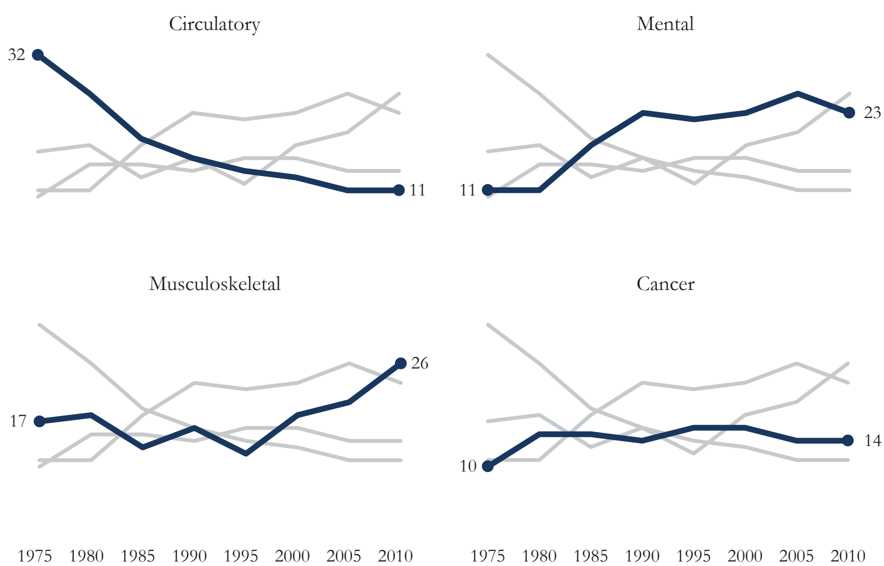


Figure 7A
A Pie Chart

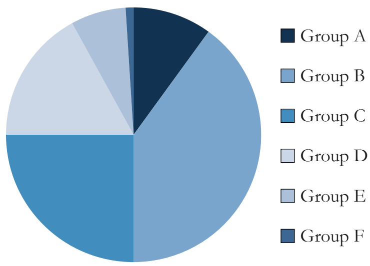
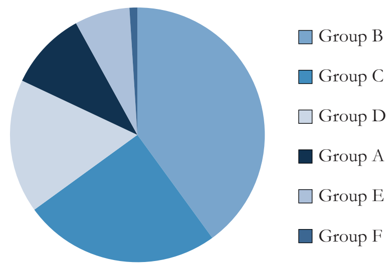


Figure 7B
B: A Pie Chart, Rotated



graphs to highlight the information in each line within the context of all the data. The contrast between light and dark helps to highlight specific trends and reduce clutter, as does the use of a label at either end of the main line in each set. (The y-axes are deleted, but they could be restored.) In this redesign, I have tried to emphasize the trends over time; my approach would differ if I were trying to emphasize specific numbers.

The Pie Chart

The debate concerning the effectiveness of pie charts is among the most contentious in the field of data visualization (much of the discussion in this section is based on Few 2007). Many people love pie charts—they are familiar, easily understood, and present "part-to-whole" relationships in an obvious way. But others argue that because pie charts force readers to make comparisons using the areas of the slices or the angles formed by the slices—something that our visual perception does not accurately support—they are not an effective way to communicate information. Donut charts—in which the center of the pie is punched out—just exacerbate the problem: The empty center makes the reader *estimate* the angle and arrive at other qualitative part-to-whole judgments without being able to see the center where the edges meet.

Pie chart slices that form 90-degree right angles—that is, slices that form one-quarter increments—are the most familiar to our eyes. Other amounts can be far more difficult to discern. The six segments in Figure 7A, for example, are presented clockwise in alphabetical order. Group C is easily identified as being about 25 percent of the whole. If, as in Figure 7B, however, the order of the segments is positioned so that the largest starts at the 12 o'clock position, the value of Group C is not so easily apprehended. One small change obscures the information.

To test your accuracy, try to guess the values of the slices: Figure 8A shows the answers. Was your guess about Group A correct? What about Group E? When I ask this question, guesses ordinarily range between 5% and 17% for each. Figure 8B offers a different approach: Add labels that integrate the data (instead of listing percentages in the notes or attaching them to the legend).

Figure 8A
A Pie Chart, Labeled

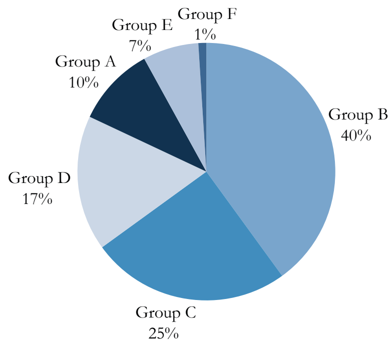
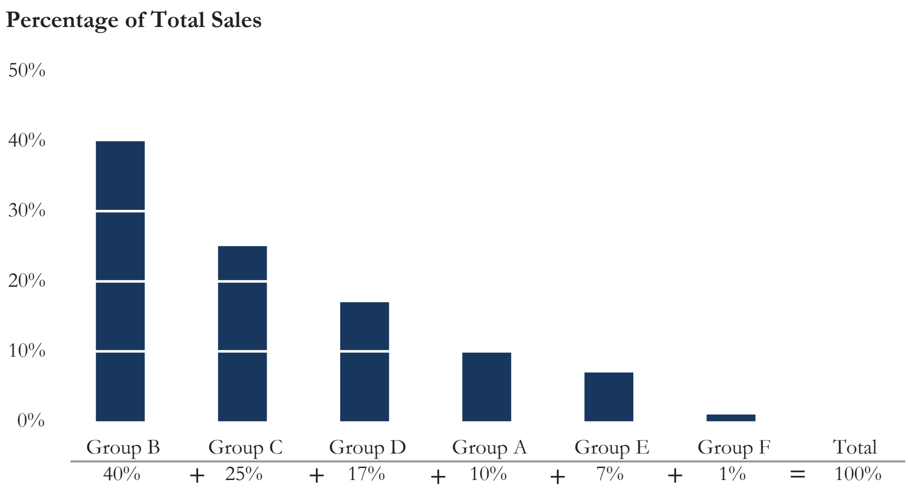
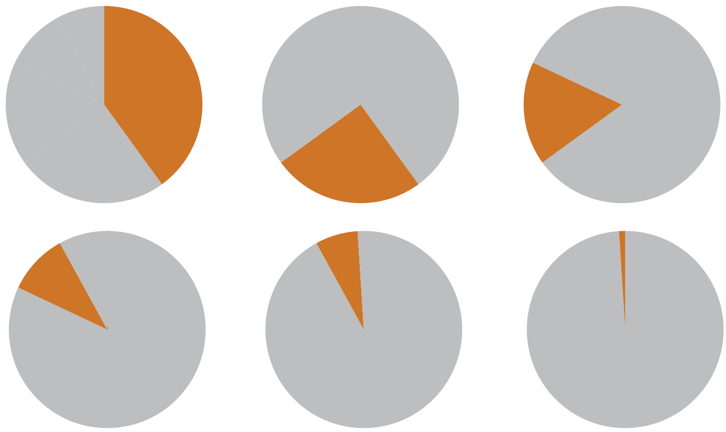


Figure 8B
Pie Chart Alternative: A Bar or Column Chart



This approach results in what amounts to two sets of information: the labels and the values for the slices (which readers might add up to arrive at 100 percent). This defeats the very purpose of the chart, which is to provide a visual representation of the data. It might be more effective just to present this information in the form of a short table. A bar or column chart could be another effective alternative: Figure 8B takes the guesswork out of identifying the value of each group and gives a clear picture of the relationships between the various groups, both in absolute amounts and in relative differences, as well as a ranking of the group from largest to smallest. The bar or column chart representation is best suited for comparing different segments; however, it is less efficient at helping us make part-to-whole

Figure 8C
Part-to-Whole Mini-Pie Charts



comparisons. In an attempt to add "part-to-whole" context—something that is often not that interesting in itself—the descriptive title at the top and the values and plus signs at the bottom reinforce the fact that the sum of the columns is 100 percent. (A completed graph could also delete the percent signs.)

Another point: Although pie charts typically ask us to compare all the data within a single space, they are by definition designed to facilitate part-to-whole comparisons. Thus, Figure 8C shows the true purpose of the pie chart, which is to individually compare each part to the whole (Camões 2013).

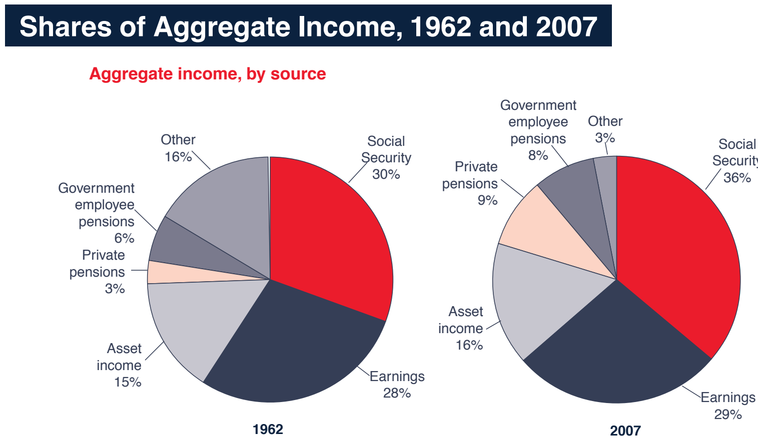
Another common—and less-than-optimal—way to present data is in a 3D pie chart. Like the 3D bar chart, such a treatment often presents the additional drawback of actually distorting the data. Whatever slice of a 3D pie chart that is toward the "front" tends to look bigger, because you can also see the 3D thickness for that slice. However, slices toward the "back" of a 3D pie chart tend to look smaller, because their "thickness" is only partially visible or not visible at all (Skau 2012).

As an example of the difficulty in discerning quantities from pie charts, Figure 9A makes the reader undertake a difficult comparison of a large number of segments—not just internally, but with another similar set right beside the first.[2]

The process of information transfer can be simplified by use of other graphic forms, which also offer the benefit of emphasizing different aspects of the analysis. For example, the *paired column chart* in Figure 9B fosters within-category comparisons. The vertical orientation requires that the labels either stretch across several lines or, worse, are rotated 90 degrees. A horizontal orientation would also work

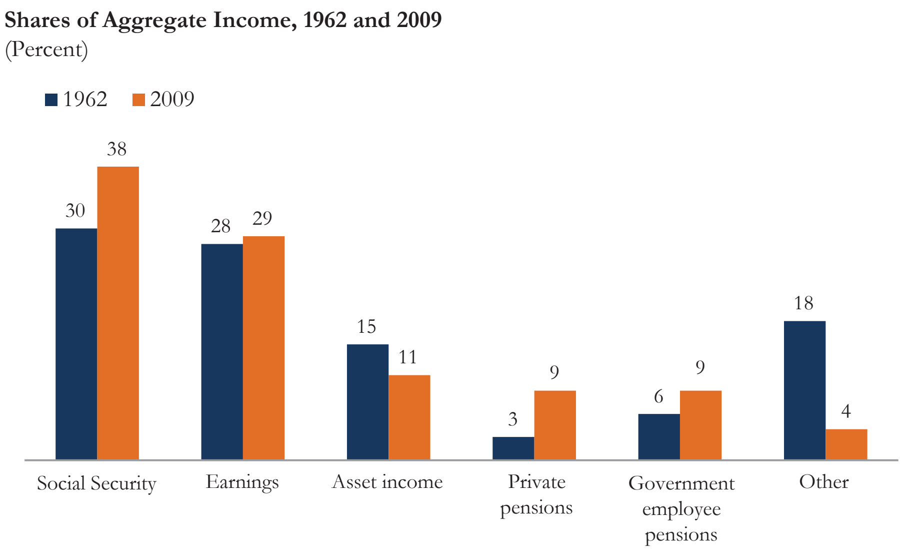
[2] Note that the pie chart for 1962 does not sum to 100 percent and there is a small, unlabeled gap left at the 12 o'clock position. For the redesigned graphs in Figures 9B–9D, I added the missing 2 percent from the 1962 series to the "Other" category. Brock (2013) shows other possible alternatives to paired pie charts.

Figure 9A
Two Pie Charts for Comparison



Source: Social Security Administration (2009).

Figure 9B
Alternative to a Pie Chart: A Paired Column Chart



in this case and is a useful approach when labels are difficult to fit in the vertical column chart layout (recall Figure 3; also see Schwabish 2013c). Also notice the (subjective) decision to omit the y-axis; the usefulness of the y-axis is doubtful with data labels placed on top of each column.

Figure 9C
Alternative to a Pie Chart: A Stacked Bar Chart

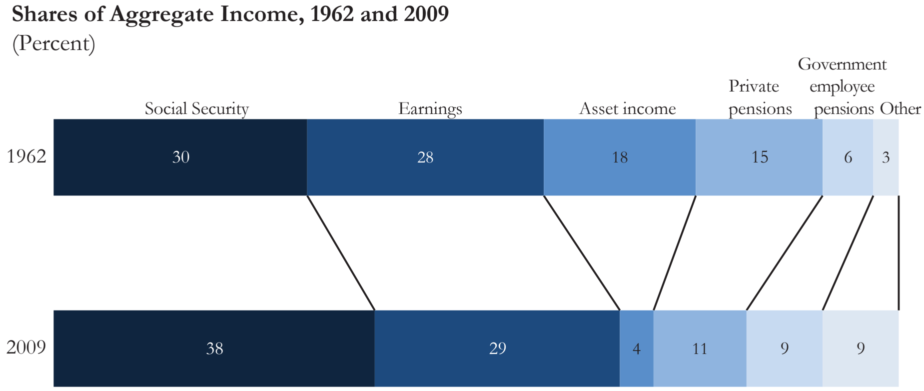
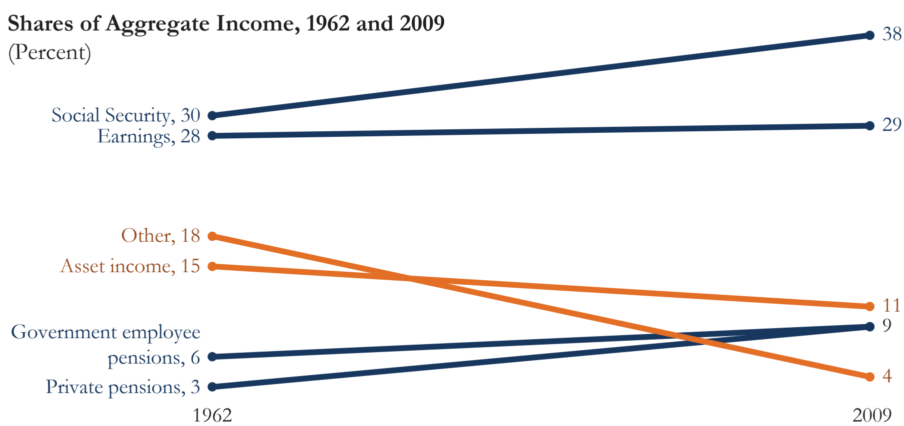
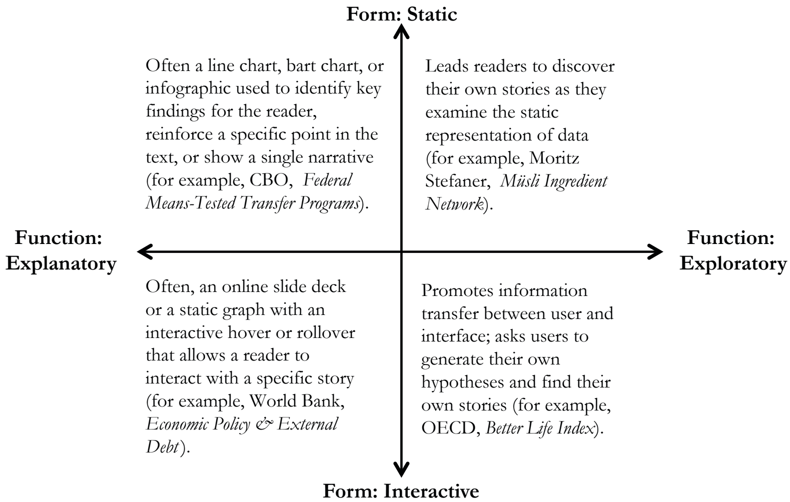


Figure 9D
Alternative to a Pie Chart: The Slope Chart



Alternatively, the *stacked bar chart* in Figure 9C shows the distribution of the various groups and that the groups sum to 100 percent, while also highlighting differences from one year to the other. Finally, the *slope chart* in Figure 9D also shows the difference in each category from the first year to the last by pairing points on two vertical axes. Slope charts can be used for a variety of purposes including showing correlations; for example, the relationship between a state's obesity rate and the share of people with at least a bachelor's degree (Cairo 2013). In this example, the color contrast (or what appears as different shades of grey in the black-and-white printed version) identifies which categories increased over time (blue; darker) and those that declined (orange; lighter).

Figure 10
Data Visualization: Form and Function



Source: CBO (Congressional Budget Office), *Means-Tested Transfer Program* (http://www.cbo.gov/publication/43935); Moritz Stefaner *Müsli Ingredient Network* (http://stefaner.eu/projects/musli-ingredient-network); World Bank, Economic Policy & External Debt (http://data.worldbank.org/topic/economic-policy-and-external-debt); and OECD, *Better Life Index* (http://www.oecdbetterlifeindex.org).

Form and Function

Once we move past the various strategies for presentation, it is useful to examine the various forms and functions of data visualization. First, consider the vertical axis in Figure 10, which illustrates the connection between the two general *forms* of visualization. *Static visualizations* provide all of the information at once and are not active or moving: for example, visualizations that appear on paper are static. *Interactive visualizations* allow a transfer of information between the figure and the user. Animated visualizations, which move but do not necessarily permit manipulation of data points to create alternative results—like movies or online slideshows where the user can control the pace of the story—can be thought of as falling between a static and an interactive visualization. Second, consider the horizontal axis in Figure 10, which considers the *function* of the visualization. *Explanatory visualizations* bring the main results to the forefront—they "surface key findings"—to some extent helping to reveal the story (for further discussion, see Segel and Heer 2010; Kosara and Mackinlay 2013). In comparison, *exploratory visualizations* help users interact with a dataset or subject matter to uncover the findings themselves. Such visualizations do not generally propose a single narrative or draw out specific insights.[3]

[3] Other mappings are of course possible and just as feasible (Bertin 1983; Harris 1996; Heer, Bostock, and Ogievetsky 2010; Kirk 2013; Kosara 2013a).

Economists typically live in a world of explanatory, static graphs. The graphs tend to be static and are used to reinforce a point made in accompanying text. Infographics (a longer form that tends to combine text, graphics, pictures, and icons) also typically are in that quadrant.

An example of a static explanatory figure would be one that portrays a single narrative, like the Congressional Budget Office's *Federal Means-Tested Transfer Programs* infographic (http://www.cbo.gov/publication/43935), which combines images, graphs, and text to tell a specific story. An example of a static exploratory figure is Moritz Stefaner's *Müsli Ingredient Network* (http://stefaner.eu/projects/musli-ingredient-network), which shows combinations of custom-ordered muesli ingredients. This approach leads readers to discover their own stories as they examine the static representation of data.

Interactive visualizations often are popular because they open the possibility of new and independent conclusions. They also enable the user to take those discoveries and produce something more explanatory. In general, the most effective interactive visualizations follow a three-step mantra: "Overview first, zoom and filter, then details-on-demand" (Shneiderman 1996). Such visualizations give a broad look at the graphic space, then allow readers to further define the space of interest, and finally permit them to capture specific details.

Perhaps the easiest explanatory interactive graph type to consider is a static graph that has an interactive hover or rollover layered on top (for example, the line graphs of economic indicators produced by the World Bank, http://data.worldbank.org/topic/economic-policy-and-external-debt). Exploratory interactive visualizations (such as OECD's *Better Life Index*, http://www.oecdbetterlifeindex.org) graphically present a complete dataset and ask users to find interesting stories. Subsequent links and reports then allow readers to find the details on demand.

Tools and Resources

Standard statistical software packages can generate basic static graphs, but people who want to improve their visualizations can take a few steps to reduce clutter and move away from default layouts, gridlines, colors, and fonts. The discussion below is not comprehensive and mentions of specific products are not intended as an endorsement but may serve as a starting point. A longer list can be found at the "Resources" page of my website (http://policyviz.com/resources/). Several tools—often free—can help users to make better use of their data in analysis or as they prepare for presentation or publication.

Color

Color could be the most misused and misunderstood aspect of design (Kosara 2013b). One simple strategy for improving the finished product is simply to avoid using default palettes. The red, green, and blue default in Excel is so pervasive that just changing to a different set of colors can make a graph more appealing.

Free online tools like Adobe Kuler (kuler.adobe.com/create/color-wheel), ColorBrewer2.0 (colorbrewer2.org), ColorSchemeDesigner (colorschemedesigner.com), and Colrd (colrd.com), allow users to create, modify, and export color palettes. The Instant Eyedropper tool (instant-eyedropper.com) allows users to choose colors from any image on a computer screen. Much has been written on color, but Ware (2012) offers perhaps the most comprehensive discussion. For situations such as publication in this journal, where one's graph may appear in both color and greyscale, there are tools—especially the ColorBrewer tool—that allow one to test grayscale-consistent color palettes.

Another important note about color is that about 10 percent of the population has some form of a color vision deficiency, and many of those people have difficulties distinguishing between greens and reds (see, for example, Coady 2013). Color Oracle (colororacle.org) and Vischeck (www.vischeck.com/) are two free tools that can simulate varieties of color blindness.

Fonts

With so many type choices on most people's computers or available online for free, it seems a shame to use the boring and overused Arial, Calibri, and Times New Roman typefaces. New fonts are designed all the time, and many old ones can give an image a fresh look. Good starting places are Font Squirrel (www.fontsquirrel.com) and Google Fonts (www.google.com/fonts). Spiekermann and Ginger (2003) and Bringhurst (2013) explain some of the science of typography and offer guides to choosing type for different purposes.

Visualization Tools

It is relatively simple to venture beyond the basic graphs available through the default settings in Stata, SAS, Excel, and other commonly available programs. The default graph in Stata, for example, has a blue background, and the first set of tick marks are not always located where the x- and y-axes intersect. Both of those conditions are easily changed. As alternatives, the open-source language R (www.r-project.org) offers more graphing capabilities. And although Excel is often dismissed as unimaginative, a variety of blogs, books, and websites provide tips and strategies to extend its capabilities (www.peltiertech.com is one).

It used to be that knowledge of HTML, JavaScript, or some other web-based programming language was a prerequisite for creating interactive visualizations. But an expanding set of available graphics tools require no more than the ability to drag and drop. Several different file types can be imported into Tableau (www.tableausoftware.com), for example, to create a variety of interactive graphics. Custom visualizations can be built using different programming languages such as HTML, JavaScript, or Processing (http://processing.org/). The R programming language also has a series of additions that enable users to create interactive visualizations: for examples, see, rCharts (http://ramnathv.github.io/rCharts/) and Shiny (http://www.rstudio.com/shiny/). The *New York Times*, for example, makes extensive use of the JavaScript library D3. Creator Mike Bostock (http://bost.ocks.org/mike/ and bl.ocks.org/mbostock) provides a library of D3 visualizations, and Murray (2013) offers an introduction to the language.

Layout

The workhorse of most graphic designers is the Adobe Creative Suite, which includes programs such as InDesign, Illustrator, and PhotoShop. The free, open-source software Inkscape (http://inkscape.org/) is an alternative. Many books have been published on the topic of layout and design: Golombisky and Hagen (2010) offer a starting point to better understanding of design techniques, and Tondreau (2009) offers a good introduction to layout.

Mapping

Formal mapping software can be quite expensive and free online versions—from ArcGIS and ESRI, for example—are often relatively inflexible. Stata offers the "spmap" add-in (www.stata.com/support/faqs/graphics/spmap-and-maps), but the quality of the images is disappointing. StatPlanet (www.statsilk.com) is a free Flash-based program that imports data from Excel to create interactive visualizations. Interactive maps can also be constructed in Tableau, mentioned above. Another free tool, TileMill (http://www.mapbox.com/tilemill/) is HTML-based and thus may require a bit more time to learn and use, but it is slightly more flexible.

Infographic Tools

Rising interest in infographics has led to the creation of services that guide users through the design process. Like those for interactive visualizations, the new infographic packages are more user friendly than older tools. Examples include Datawrapper (http://datawrapper.de/), Infogr.am (http://infogr.am/), and Lemon.ly (http://lemon.ly/).

Resources

The amount of writing on data visualization has exploded over the past few years. The data visualization field moves fast—new work is constantly emerging, new products are constantly being released, and discussion and debates about best practices are continuous. Many books and blogs offer in-depth discussions on data visualization techniques and strategies or offer tutorials for the tools listed above.

Early and fundamental books are those by Tukey (1977), Bertin (1983), Cleveland (1993), and Tufte (2001[1983]). Wong (2010) and Robbins (2013b) are excellent sources for classifying specific charts for specific data types. In his books, Few (2009, 2012) dedicates a bit more time to examine the cognitive theory of data visualization and effective data visualization techniques. Recent books by Cairo (2013) and Yau (2011, 2013) are also excellent newer contributions to the field.

The number of blogs dedicated to the data visualization field is constantly growing. What follows is a very short list: Eager Eyes (eagereyes.org) is produced by Robert Kosara, a visual analysis researcher at Tableau Software and former computer science professor at UNC-Charlotte. Kosara often writes about the research side of data and information visualization and offers constructive criticism and explorations. Flowing Data (flowingdata.com) is produced by author and statistician Nathan Yau, who provides a daily showcase of visualizations from around the web. He also posts visualization tutorials, primarily in the R programming language. Perceptual Edge (perceptualedge.com) is produced by author and consultant Stephen Few, who writes about good and bad trends in data visualization and promotes best practices with a focus on the visual aspects of human cognition. Junk Charts (junkcharts.typepad.com) is where Kaiser Fung collects and offers criticism of charts that do a poor job of presenting data. Visualising Data (visualisingdata.com) author Andy Kirk details the design process and trends in the field. Storytelling with Data (storytellingwithdata.com) is Cole Nussbaumer's blog where she offers practical examples for creating more effective visualizations. Finally, I offer practical data visualization examples on my website (policyviz.com/), as well as issues pertaining to creating more effective verbal presentations. On my companion site, HelpMeViz.com, readers can submit works in progress to seek advice and feedback from the data visualization community.

Conclusion

For economists who want readers to apprehend results quickly and accurately, presentation matters. Effective visualizations show the data to tell the story, reduce clutter to keep the focus on the important points, and integrate the text with the graphs to transfer information efficiently. With the increased flexibility of even fairly basic software programs (like Excel), it is now more cost-effective in terms of time and energy for researchers to invest some time learning and thinking about the details of graphical presentation.

To create great, effective visualizations, carefully consider the needs of your audience—the numbers, facts, or stories that will help them understand your ideas and your arguments. Consider the interfaces—static versus interactive—they will use. And pair the depth and clarity of your data, models, and writing with visualizations that are just as clear and compelling.

■ *The views in this article are mine and should not be interpreted as those of the Congressional Budget Office. For more information on data visualization and presentation techniques, see my site www.PolicyViz.com. I am grateful to the editors of this journal—David Autor, Chang-Tai Hseih, Ulrike Malmendier, and Timothy Taylor—for recognizing the importance of data visualization in the economics profession, and to Ann Norman and Annette Blanar for their painstaking proofreading and typesetting work. I also thank the authors of the graphs used in the paper and Alberto Cairo, Stephen Few, Molly Dahl, Kate Kelly, and Robert Kosara.*

References

Bertin, Jacques. 1983. *Semiology of Graphics: Diagrams, Networks, Maps.* New York, NY: Esri Press.

Bringhurst, Robert. 2013. *Elements of Typographic Style: Version 4.0, 20th Anniversary Edition.* Point Roberts, WA: Hartley and Marks.

Brock, Tim. 2013. "Improving on a Pair of Pie Charts." http://datatodisplay.com/blog/chart-design/improving-pair-pie-charts/.

Cairo, Alberto. 2013. *The Functional Art: An Introduction to Information Graphics and Visualization.* Berkeley, CA: New Riders.

Camões, Jorge. 2013. "Finally Revealed: The Optimal Number of Categories in a Pie Chart." http://www.excelcharts.com/blog/optimal-number-categories-pie-chart/.

Cleveland, William S. 1993. *Visualizing Data.* Summit, NJ: Hobart Press.

Coady, Geri. 2013. *Colour Accessibility.* (The Pocket Guide series: Collection 3.) Five Simple Steps. http://www.fivesimplesteps.com/products/the-pocket-guide-series-collection-three.

Corum, Jonathan. 2013. "Storytelling with Data." Opening keynote address at the Tapestry Conference, held in Nashville, TN, February 23. http://style.org/tapestry/.

Few, Stephen. 2004. "Tapping the Power of Visual Perception." *Visual Business Intelligence Newsletter,* September 4. http://www.perceptualedge.com/articles/ie/visual_perception.pdf.

Few, Stephen. 2005. "Grid Lines in Graphs are Rarely Useful." *Visual Business Intelligence Newsletter,* February. http://www.perceptualedge.com/articles/dmreview/grid_lines.pdf.

Few, Stephen. 2007. "Save the Pies for Dessert." *Visual Business Intelligence Newsletter* (August).

Few, Stephen. 2009. *Now You See It.* Burlingame, CA: Analytics Press.

Few, Stephen. 2012. *Show Me the Numbers: Designing Tables and Graphs to Enlighten.* 2nd ed. Burlingame, CA: Analytics Press.

Golombisky, Kim, and Rebecca Hagen. 2010. *White Space is Not Your Enemy: A Beginner's Guide to Communicating Visually through Graphic, Web & Multimedia Design.* Burlington, MA: Elsevier, Inc.

Hanson, Gordon H. 2012. "The Rise of Middle Kingdoms: Emerging Economies in Global Trade." *Journal of Economic Perspectives* 26(2): 41–64.

Harris, Robert L. 1996. *Information Graphics: A Comprehensive Illustrated Reference.* New York, NY: Oxford University Press.

Healey, Christopher G., and James T. Enns. 2012. "Attention and Visual Memory in Visualization and Computer Graphics." *IEEE Transactions on Visualization and Computer Graphics* 18(7): 1170–88.

Heer, Jeffrey, Michael Bostock, and Vadim Ogievetsky. 2010. "A Tour through the Visualization Zoo." ACMQueue, May 13. http://queue.acm.org/detail.cfm?id=1805128.

Hockley, William E., and Tyler Bancroft. 2011. "Extensions of the Picture Superiority Effect in Associative Recognition." *Canadian Journal of Experimental Psychology* 65(4): 236–56.

Kirk, Andy. 2013. "Discussion: Storytelling and Success Stories." April. http://www.visualisingdata.com/index.php/2013/04/discussion-storytelling-and-success-stories/.

Klerman, Jacob Alex, and Caroline Danielson. 2011. "The Transformation of the Supplemental Nutrition Assistance Program." *Journal of Policy Analysis and Management* 30(4): 863–88.

Kosara, Robert. 2013a. "Visual Storytelling in the Age of Data." Keynote address at the Tapestry Conference, held February in Nashville, TN. http://www.youtube.com/watch?v=qSYEjhR2AwQ&feature=youtu.be.

Kosara, Robert. 2013b. "How the Rainbow Color Map Misleads." July 7. http://eagereyes.org/basics/rainbow-color-map.

Kosara, Robert, and Jock Mackinlay. 2013. "Storytelling: The Next Step for Visualization." *Computer (Special Issue on Cutting-Edge Research in Visualization)* 46(5): 44–50.

Medina, John. 2008. *Brain Rules: 12 Principles for Surviving and Thriving at Work, Home, and School.* Pear Press.

Murray, Scott. 2013. *Interactive Data Visualization for the Web: An Introduction to Designing with D3.* Sebastopol, CA: O'Reilly Media, Inc.

Nussbaumer, Cole. 2013. "Strategies for Avoiding the Spaghetti Graph." March 14. http://www.storytellingwithdata.com/2013/03/avoiding-spaghetti-graph.html.

Organisation for Economic Co-operation and Development. Accessed, August 2013. "Percentage of Employed Who are Senior Managers, by Sex." http://www.oecd.org/gender/data/proportionofemployedwhoareseniormanagersbysex.htm.

Ottaviano, Gianmarco I. P., and Giovanni Peri. 2008. "Immigration and National Wages: Clarifying the Theory and the Empirics." NBER Working Paper 14188 (July).

Rampell, Catherine. 2013. "Comparing the World's Glass Ceilings." *Economix,* April 2. http://economix.blogs.nytimes.com/2013/04/02/comparing-the-worlds-glass-ceilings/?_r=0.

Robbins, Naomi. 2013a. "How to Position Y-Axis Labels in Graphs." February 12. http://www.forbes.com/sites/naomirobbins/2013/02/12/how-to-position-y-axis-labels-in-graphs/.

Robbins, Naomi. 2013b. *Creating More Effective Graphs*. 2nd ed. Chart House.

Schwabish, Jonathan. 2013a. "Mind the Gap—An Economic Remake." The Why Axis, April. http://thewhyaxis.info/gap-remake/.

Schwabish, Jonathan. 2013b. "To Label or Not to Label? That *Is* the Question." June 10. http://www.allanalytics.com/author.asp?section_id=3072&doc_id=264322.

Schwabish, Jonathan. 2013c. "Visualizing Data: Bad Labels, Easy Fix." May 2. http://www.allanalytics.com/author.asp?section_id=3072&doc_id=262539.

Segel, Edward, and Jeffrey Heer. 2010. "Narrative Visualization: Telling Stories with Data." *IEEE Transactions on Visualization and Computer Graphics* 16(6): 1139–48.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." In *Proceedings of the IEEE Symposium on Visual Languages*, IEEE Computer Society Press, 336–43.

Skau, Drew. 2012. "2D's Company, 3D's a Crowd." http://blog.visual.ly/2ds-company-3ds-a-crowd/.

Social Security Administration. 2009. *Fast Facts & Figures About Social Security, 2009.* Social Security Administration (July). http://www.ssa.gov/policy/docs/chartbooks/fast_facts/2009/fast_facts09.pdf.

Social Security Advisory Board. 2012. *Aspects of Disability Decision Making: Data and Materials.* Social Security Administration (February), http://www.ssab.gov/Publications/Disability/GPO_Chartbook_FINAL_06122012.pdf.

Spiekermann, Erik, and E. M. Ginger. 2003. *Stop Stealing Sheep & Find Out How Type Works,* 2nd ed. Berkeley, CA: Adobe Press.

Stinebrickner, Ralph, and Todd Stinebrickner. 2013. "Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model." *Western University CIBC Working Paper 2013-5* (July), http://economics.uwo.ca/cibc/workingpapers_docs/wp2013/Stinebrickner_Stinebrickner05.pdf.

Tondreau, Beth. 2009. *Layout Essentials: 100 Design Principles for Using Grids*. Beverly, MA: Rockport Publishers.

Tukey, John W. 1977. *Exploratory Data Analysis.* Reading, MA: Addison Wesley.

Tufte, Edward. 2001 [1983]. *The Visual Display of Quantitative Information*. 2nd ed. (First edition 1983). Cheshire, CT: Graphics Press.

Tufte, Edward. 2006. *Beautiful Evidence.* Cheshire, CT: Graphics Press.

Ware, Colin. 2012. *Information Visualization: Perception for Design*. 3rd ed. Waltham, MA: Morgan Kaufman.

Wong, Dona M. 2010. *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures.* New York: W. W. Norton and Company.

Yau, Nathan. 2011. *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics.* Indianapolis, IN: Wiley Publishing, Inc.

Yau, Nathan. 2013. *Data Points: Visualization that Means Something*. Indianapolis, IN: Wiley Publishing, Inc.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 235–242

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.

Smorgasbord

An IMF staff team with Bernardin Akitoby as lead author considers how conventional wisdom about fiscal policy has changed in "Reassessing the Role and Modalities of Fiscal Policy in Advanced Economies." "The prevailing consensus before the crisis was that discretionary fiscal policy had a limited role to play in fighting recessions. The focus of fiscal policy in advanced economies was often on the achievement of medium- to long-run goals such as raising national saving, external rebalancing, and maintaining long-run fiscal and debt sustainability given looming demographic spending pressures. For the management of business cycle fluctuations, monetary policy was seen as the central macroeconomic policy tool. Fiscal contraction was sometimes recommended during periods of economic overheating as a means of supporting monetary policy, for example to take pressure

■ *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.28.1.235.   doi=10.1257/jep.28.1.235

off the exchange rate in the face of persistent capital inflows. However, during downturns, it was deemed that there was little reason to use another instrument beyond monetary policy. . . . The crisis has provided evidence that fiscal policy is an appropriate countercyclical policy tool when monetary policy is constrained by the zero lower bound, the financial sector is weak, or the output gap is particularly large. Nevertheless, a number of reservations regarding the use of discretionary fiscal policy tools remain valid, particularly when facing 'normal' cyclical fluctuations." IMF Policy Paper, September 2013, http://www.imf.org/external/np/pp/eng/2013/072113.pdf.

Deborah K. Elms and Patrick Low have edited an e-book, *Global Value Chains in a Changing World.* The 16 chapters look at topics like how to measure the value-added within each country, how to manage these production processes, and how low- and medium-income countries can find a niche in these production chains. Richard Baldwin contributed the overview essay. "Global supply chains have transformed the world. They revolutionized development options facing poor nations; now they can join supply chains rather than having to invest decades in building their own. The offshoring of labour-intensive manufacturing stages and the attendant international mobility of technology launched era-defining growth in emerging markets, a change that fosters and is fostered by domestic policy reform. This reversal of fortunes constitutes perhaps the most momentous global economic change in the last 100 years. Global supply chains, however, are themselves rapidly evolving. The change is in part due to their own impact (income and wage convergence) and in part due to rapid technological innovations in communication technology, computer integrated manufacturing and 3D printing." Baldwin argues that the "first great unbundling" of international trade tended to cluster economic production in certain regions of the global North. In contrast, he argues that the "second great unbundling" of global supply chains will spread economic growth more broadly. As one measure of these shifting patterns, Baldwin points out that the G-7 economies—the United States, Canada, France, Germany, Italy, Japan, and the United Kingdom—represented 20 percent of global output in 1820, 40 percent of global output in 1870, and peaked at two-thirds of global output in 1988, but have now fallen back to 50 percent of global output. Published by the World Trade Organization, together with the Fung Global Institute and Nanyang Technological University, 2013. http://www.wto.org/english/res_e/booksp_e/aid4tradeglobalvalue13_e.pdf.

Robert Z. Lawrence and Lawrence Edwards discuss "US Employment Deindustrialization: Insights from History and the International Experience." "We argue that while trade has contributed to the decline in US manufacturing employment, many exaggerate the role of international competition. The US manufacturing employment decline results less from international factors and is instead driven by powerful historical forces that have affected all advanced economies. These are a combination of rapid productivity growth and demand that is relatively unresponsive to income growth and lower prices. We do not claim that international factors do not affect manufacturing. Our estimates suggest that the labor content of the US manufacturing trade deficit remains significant and that despite improvements in US competitiveness, a vigorous US and global economic recovery could boost

US manufacturing employment. Over the long run, however, absent new product innovations, or a shift in consumer preferences, the basic forces leading to the declining share of manufacturing in overall employment are unlikely to abate. Just as rapid farm productivity growth combined with a limited demand for food has led to ever smaller shares of employment in agriculture, the combination of relatively rapid productivity growth and limited demand growth for goods will mean that more of the jobs in the future will be in services." Peterson Institute for International Economics, Policy Brief PB13-27, October 2013, http://www.iie.com/publications/pb/pb13-27.pdf. This paper can be read as a useful accompaniment to the three-paper symposium on manufacturing in this issue.

The Global CCS Institute discusses issues facing carbon capture and storage (CCS) technology in its annual report, *The Global Status of CCS: 2013*. "CCS has strong potential to be cost competitive in a low–carbon future. The International Energy Agency (IEA) has estimated that the exclusion of CCS as a technology option in the electricity sector alone would increase mitigation costs by around US$2 trillion by 2050." However, the report also details many of the challenges. "More than 90 per cent of the overall cost of CCS can be driven by expenses related to the capture process. . . . There is a variety of R&D programs focused on developing new and more cost-effective capture technologies." "For CCS to meet the longer term climate challenge of restricting global warming to less than 2°C, the estimated magnitude of the CO_2 transportation infrastructure that will need to be built in the coming 30–40 years is 100 times larger than currently operating CO_2 pipeline networks." "[T]here is an urgent need for policies and funded programs that encourage the exploration and appraisal of significant CO_2 storage capacity." At http://cdn.globalccsinstitute.com/sites/default/files/publications/116211/global-status-ccs-2013.pdf.

Anthony P. Carnevale, Andrew R. Hanson, and Artem Gulish discuss *Failure to Launch: Structural Shift and the New Lost Generation*. "Young adults do not reach levels of employment and earnings levels similar to those of young adults in 1980 until later ages, while older workers are employed at higher rates and have consistently higher relative earnings than older workers from a generation ago. The age at which young adults reach the median wage shifted from 26 to 30 between 1980 and 2012. . . . In other words, the model of the labor market that presumed entry at age 18 and exit at age 65 is obsolete, and instead, young people often start their careers later, after developing more human capital from postsecondary education and training, and work experience from internships, work-study, mentorships, fellowships, job shadowing, and part-time work. Young people today change jobs more frequently between the ages of 18 and 25 and only one out of 10 describes his or her current job as a career. . . . Evolving social norms entangled with economic hardships have led young people to delay household and family formation. Two-thirds of young adults in their 20s cohabitate; the average age of marriage increased from 21 to 26 for women and 23 to 28 for men between 1970 and 2006. Over the same period, the average age of a mother at the birth of her first child increased from 21 to 25. In 1960, three out of four women and two out of three men completed school, left home, achieved financial independence, were married and had children by age 30. In 2000, less than half of women and only one-third of men reached the same milestones by age 30." Georgetown Public Policy Institute, September 2013, http://www9.georgetown.edu/grad/gppi/hpi/cew/pdfs/FTL_FullReport.pdf.

Richard Dobbs, Susan Lund, Tim Koller, and Ari Shwayder analyze "QE and Ultra-Low Interest Rates: Distributional Effects and Risks." They acknowledge and accept the conventional wisdom that the ultra-low interest rate policies were useful and appropriate as part of the effort to stave off the Great Recession, though controversy remains over continuing the policies. They tackle instead the distributional questions about ultra-low interest rates. "From 2007 to 2012, governments in the eurozone, the United Kingdom, and the United States collectively benefited by $1.6 trillion both through reduced debt-service costs and increased profits remitted from central banks (exhibit). Nonfinancial corporations—large borrowers such as governments—benefited by $710 billion as the interest rates on debt fell. Although ultra-low interest rates boosted corporate profits in the United Kingdom and the United States by 5 percent in 2012, this has not translated into higher investment, possibly as a result of uncertainty about the strength of the economic recovery, as well as tighter lending standards. Meanwhile, households in these countries together lost $630 billion in net interest income, although the impact varies across groups. Younger households that are net borrowers have benefited, while older households with significant interest-bearing assets have lost income." McKinsey Global Institute Discussion Paper, November 2013. http://www.mckinsey.com/insights/economic_studies/qe_and_ultra_low_interest_rates_distributional_effects_and_risks?cid=other-eml-alt-mgi-mck-oth-1311.

The Credit Suisse Research Institute ponders "Sugar Consumption at a Crossroads." "The global sweetener market is estimated to be around 190 million tons of 'white sugar equivalent,' and is unsurprisingly dominated by sugar. Each of the major groups (high-intensity/artificial sweeteners, sugar, and high-fructose corn syrup) has been growing at a similar rate of circa 2% per annum, though the most recent numbers have natural high-intensity sweeteners growing rather faster." "Sweetened beverages are now delivering an increasingly greater percentage of the sugars that are ingested in an average diet. Between 1955 and 2000, the consumption of soft drinks in the USA increased from about ten gallons/person to 54 gallons/person and then declined by around 20% until 2012, but with an equivalent increase in the consumption of fruit juices and bottled water. According to the USDA, the beverage industry now accounts for 31% of total sweetener deliveries and we estimate that 43% of added sugars in a normal US diet come from sweetened beverages. A similar stabilizing trend can be seen in most other developed markets, while consumption is still on the rise in emerging markets." The report also looks at evidence linking sugar consumption in various forms to obesity, the health consequences of obesity, and possible policy options. September 2013, https://publications.credit-suisse.com/tasks/render/file/index.cfm?fileid=780BF4A8-B3D1-13A0-D2514E21EFFB0479.

The US Department of Education has published "Literacy, Numeracy, and Problem Solving in Technology-Rich Environments Among U.S. Adults: Results from the Program for the International Assessment of Adult Competencies 2012: First Look," written by Madeline Goodman, Robert Finnegan, Leyla Mohadjer, Tom

Krenzke, and Jacquie Hogansee. The PIAAC is a test given to nationally representative samples of adults in 23 high-income countries in 2011–2012. On the literacy scale: "Compared with the U.S. average score, average scores in 12 countries were higher, in 5 countries they were lower, and in 5 countries they were not significantly different." On the numeracy scale: "Compared with the U.S. average score, average scores in 18 countries were higher, in 2 countries they were lower, and in 2 countries they were not significantly different." On scores for "problem solving in technology-rich environments": "Compared with the U.S. average score, average scores in 14 countries were higher and in 4 countries they were not significantly different." At https://nces.ed.gov/pubs2014/2014008.pdf. See also the related OECD reports: *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*, available at http://skills.oecd.org/documents/OECD_Skills_Outlook_2013.pdf, and "Time for the U.S. to Reskill? What the Survey of Adult Skills Says," at http://skills.oecd.org/Survey_of_Adult_Skills_US.pdf.

Homelessness in America has declined in recent years, even in the aftermath of the Great Recession. The US Department of Housing and Urban Development has published "The 2013 Annual Homeless Assessment Report (AHAR) to Congress." "In January 2013, 610,042 people were homeless on a given night. Most (65 percent) were living in emergency shelters or transitional housing programs and 35 percent were living in unsheltered locations. . . . Homelessness declined by nearly 4 percent (or 23,740 people) between 2012 and 2013, and by 9 percent (or 61,846) since 2007." At https://www.onecpd.info/resources/documents/AHAR-2013-Part1.pdf. For additional discussion of trend in US homelessness with some policy analysis, see the April 2013 report of the National Alliance to End Homelessness, "The State of Homelessness in America 2013," at http://b.3cdn.net/naeh/bb34a7e4cd84ee985c_3vm6r7cjh.pdf.

From the Federal Reserve Banks

Douglas Clement has an interview with Richard Thaler. Here's Thaler remembering the roots of his interest in what has come to be called "behavioral economics": "[L]ater I would call them anomalies, but for a while I just called them 'the list.' And I started writing a list of funny behaviors on my blackboard, such as paying attention to sunk costs. I mean, at first they were just stories. Like, a buddy of mine and I were given tickets to a basketball game. Then there's a blizzard and we don't go. But he says, 'If we had paid for the tickets, we would have gone.' Another thing on the list was a story about having a group of fellow grad students over for dinner and putting out a large bowl of cashew nuts. We started devouring them. After a while, I hid the bowl in the kitchen and everyone thanked me. But as econ grad students, of course, we immediately started asking why we were happy about having a choice removed. For years, some of my friends referred to my new research interests as 'cashew theory.'" "The biggest surprise about behavioral economics, I think, looking back on it all, is that the subfield where behavioral has had the biggest impact is finance. If you had asked me in 1980 to say which field do you think you have your best shot at affecting, finance would have been the least likely, essentially because of the arguments that [Gary] Becker's making: The stakes are really high, and you don't survive very long if you're a trader who loses money. But for me, of course, that was exactly the attraction of studying finance . . ." *The Region*: Federal Reserve Bank of Minneapolis, September 2013, pp. 14–28. http://www.minneapolisfed.org/publications_papers/pub_display.cfm?id=5184.

In an "Interview" conducted by Jessie Romero, John Haltiwanger discusses changing patterns of job creation and destruction: "But now we're seeing a decline in the entry rate and a pretty stark decline in the share of young businesses. . . . But it's also important to recognize that the decline in the share of young firms has occurred because the impact of entry is not just at the point of entry, it's also over the next five or 10 years. A wave of entrants come in, and some of them grow very rapidly, and some of them fail. That dynamic has slowed down. . . . If you look at young small businesses, or just young businesses period, the 90th percentile growth rate is incredibly high. Young businesses not only are volatile, but their growth rates also are tremendously skewed. It's rare to have a young business take off, but those that do add lots of jobs and contribute a lot to productivity growth. We have found that startups together with high-growth firms, which are disproportionately young, account for roughly 70 percent of overall job creation in the United States. . . . "I think the evidence is overwhelming that countries have tried to stifle the [job] destruction process and this has caused problems. I'm hardly a fan of job destruction per se, but making it difficult for firms to contract, through restricting shutdowns, bankruptcies, layoffs, etc., can have adverse consequences. The reason is that there's so much heterogeneity in productivity across businesses. So if you stifle that destruction margin, you're going to keep lots of low-productivity businesses in existence, and that could lead to a sluggish economy. I just don't think we have any choice in a modern market economy but to allow for that reallocation to go on. Of course, what you want is an environment where not only is there a lot of job destruction, but also a lot of job creation, so that when workers lose their jobs they either immediately transit to another job or their unemployment duration is low." *Econ Focus*, Federal Reserve Bank of Richmond, Second Quarter 2013, pp. 30–34. http://www.richmondfed.org/publications/research/econ_focus/2013/q2/pdf/interview.pdf.

Michael T. Owyang, E. Katarina Vermann, and Tatevik Sekhposyan, re-examine Okun's law in "Output and Unemployment: How Do They Relate Today?" "In his 1962 paper, Okun used data on the quarter-to-quarter growth rate of the real gross national product (GNP) and the quarter-to-quarter difference in the unemployment rate from 1947 to 1960. He estimated that if real GNP growth were held at zero, the unemployment rate would grow 0.3 percentage points, on average, from one quarter to the next. In addition, for each 1-percentage-point increase in real GNP growth, the unemployment rate would decrease 0.3 percentage points. Economists call this latter number Okun's coefficient. This empirical relationship—dubbed Okun's law—has remained largely intact for 50 years." *Regional Economist*, Federal Reserve Bank of St. Louis, October 2013, pp. 4–9. http://www.stlouisfed.org/publications/re/articles/?id=2421.

Charles Davidson describes "The Big Busy: A Radical Reset after the Katrina Catastrophe is Transforming the Economy of New Orleans." "For various reasons, the economy in New Orleans since Katrina has reversed decades of decline and outperformed the nation and other southeastern metropolitan areas. The NOLA—locals' preferred nickname—metropolitan statistical area's real gross domestic product (GDP) grew an average 3.9 percent a year from 2008 through 2011 . . . During the same period, U.S. GDP expanded less than 1 percent a year, and no other southeastern metro area topped 2 percent. Undoubtedly, the more than $100 billion in federal aid for post-Katrina rebuilding helped cushion New Orleans and south Louisiana from the Great Recession. Beyond that, though, tourism hit record levels last year, knowledge-based industries such as higher education and engineering are expanding, locals have started companies at a per capita rate 56 percent above the national average, wages have risen faster than elsewhere in the United States, and the demographic mix has shifted toward more higher-income households and fewer poor households. Finally, the New Orleans metro population as of mid-2012 had climbed back to 1.23 million, about 94 percent of its pre-Katrina level, according to the U.S. Census Bureau. In many respects, economic data describe New Orleans as a city catching up to the nation and other metro areas. . . . The arrival of more college graduates may have plugged a 'brain drain' that long plagued New Orleans . . ." *Econ Focus*, Federal Reserve Bank of Richmond, Second Quarter 2013, pp. 30–34. http://www.frbatlanta.org/documents/pubs/econsouth/13q3_big_busy.pdf. This article can be read as a follow-up to the article by Jacob Vigdor, "The Economic Aftermath of Hurricane Katrina," in the Fall 2008 issue of this journal (pp. 135–54).

Conversation Starters

Chris Edwards makes the case for "Privatizing the Transportation Security Administration." "More than 80 percent of Europe's commercial airports use private screening companies, including those in Britain, France, Germany, and Spain. The other airports in Europe use their own in-house security, but no major country in Europe uses the national government's aviation bureaucracy for screening. Europe's airports moved to private contracting during the 1980s and 1990s after numerous hijackings and terrorist threats, and it has worked very well. Canada also uses private screening companies at its commercial airports, and some airports also use private firms for general airport security. . . . The 2001 legislation that created TSA established the SPP [Screening Partnership Program], which has allowed some airports to opt out of TSA screening and use private firms. The firms contract with TSA and are under federal regulatory control. Originally, there were five airports in the program, with San Francisco being the largest. All five have had good results with private screening and have stuck with it. The number of SPP airports has grown to 16 today." Cato Institute Policy Analysis No. 742, November 19, 2013, http://www.cato.org/publications/policy-analysis/privatizing-transportation-security-administration.

Andrzej Rapaczynski discusses "The Moral Significance of Economic Life" in the most recent issue of *Capitalism and Society*. His abstract summarizes the argument compactly: "Much of the modern perception of the role of economic production in human life—whether on the Left or on the Right of the political spectrum—views it as an inferior, instrumental activity oriented toward self-preservation, self-interest, or profit, and thus as essentially distinct from the truly human action concerned with moral values, justice, and various forms of self-fulfillment. This widely shared worldview is rooted, on the one hand, in the Aristotelian tradition that sees labor as a badge of slavery, and freedom as lying in the domain of politics and pure (not technical) knowledge, and, on the other hand, in the aristocratic mediaeval Christian outlook, which—partly under Aristotle's influence—sees nature as always already adapted (by divine design) to serving human bodily needs, and the purpose of life as directed toward higher, spiritual reality. . . . As against this, liberal thinkers, above all Locke, have developed an elaborate alternative to the Aristotelian worldview, reinterpreting the production process as a moral activity par excellence consisting in a gradual transformation of the alien nature into a genuinely human environment reflecting human design and providing the basis of human autonomy. Adam Smith completed Locke's thought by explaining how production is essentially a form of cooperation among free individuals whose self-interested labor serves the best interest of all. The greatest "culture war" in history is to re-establish the moral significance of economic activity in the consciousness of modern political and cultural elites." *Capitalism and Society*, December 2013, vol. 8, no. 2, http://capitalism.columbia.edu/volume-8-issue-2.

Ted Gayer and Emily Parker discuss "Cash for Clunkers: An Evaluation of the Car Allowance Rebate System." "The Car Allowance Rebate System (CARS), more commonly known as 'Cash for Clunkers' . . . allowed consumers to trade in an older, less fuel-efficient vehicle for a voucher to be applied toward the purchase of a newer, more fuel-efficient vehicle. Depending on the difference in fuel economy between the trade-in vehicle and the new vehicle, program participants received a voucher for either $3,500 or $4,500. After the "clunker" was traded in at the dealership, its engine was destroyed, ensuring its permanent removal from the U.S. vehicle fleet. Nearly 700,000 clunkers were traded in between July 1, 2009 and August 24, 2009 under the program. . . . Our evaluation of the evidence suggests that the $2.85 billion in vouchers provided by the program had a small and short-lived impact on gross domestic product, essentially shifting roughly a few billion dollars forward from the subsequent two quarters following the program. The implied cost per job created due to the program was much higher than what was estimated for alternative fiscal stimulus programs. . . . On the environmental side, the cost per ton of carbon dioxide reduced due to the program was higher than what would be achieved through a more cost-effective policy such as a carbon tax or cap-and-trade, but was comparable (or indeed lower) than what is achieved through some of the less cost-effective environmental policies, such as the tax subsidy for electric vehicles." Brookings Institution Policy Brief, October 31, 2013. http://www.brookings.edu/research/papers/2013/10/cash-for-clunkers-evaluation-gayer.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 243–248

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 One Percent

The cheerful blandness of N. Gregory Mankiw's "Defending the One Percent" (Summer 2013, pp. 21–34) may divert attention from its occasional unstated premises, dubious assumptions, and omitted facts. I have room to point only to a few such weaknesses; but the One Percent are pretty good at defending themselves, so that any assistance they get from the sidelines deserves scrutiny.

First, the paper starts off by invoking an iconic Steve Jobs and, without making an explicit claim, carries on tacitly as if the One Percent consists mainly of entrepreneurs whose innovations generate a lot of consumer surplus for the world. There would be less alarm if that were so. But a large and—before the crisis—increasing part of the income of the One Percent arises in the financial services industry, and a large and increasing part of that has come from trading profits. Much of the income in question must therefore be the payoff to asymmetric information, and generates precious little in the way of aggregate consumer surplus. Consider this history (for details see Murphy 2012, especially pp. 307–8): from 1970 to about 1995, the median realized compensation for chief executive officers in Standard and Poor's 500 broker-dealer firms was essentially indistinguishable from that of Standard and Poor's 500 banks and industrials. Rather suddenly, between 1996 and 2006, the median broker-dealer chief executive officer started to collect anywhere between 7 and 10 times the median compensation of the other two groups. This does not smell like the Goldin and Katz (2008) story. I'll swallow "innovation," but socially productive innovation, no thanks. Extreme inequality is not primarily about useful entrepreneurs. On financial profits and inequality, Mankiw waffles uncomfortably.

Second, Mankiw refers at best tangentially to what may be the most dangerous adverse consequence of extreme inequality at the top: the rich, with a large assist from the *Citizens United v. Federal Election Commission* (558 US 310 [2010]) decision, can buy political influence, and not only influence but power. An immediate example is the millions of dollars spent by the financial services industry to weaken regulation under the Dodd–Frank Wall Street Reform and Consumer Protection Act of 2010. Mankiw's partial reference is in connection with the contribution of rent-seeking to inequality. There his response is that the proper remedy is not to attack extreme inequality but to suppress rent-seeking. Presumably he would give the same answer here: if inequality corrupts politics, go after the corruption, not the inequality. But this would be, willfully or not, naive: it is precisely the power of great wealth that makes it difficult or impossible to eradicate corruption.

Third, Mankiw's conception of rents is too narrow; he refers mainly to monopoly rents. But economic rents are pervasive. He mentions the close correlation between height and income. It seems unlikely that height is correlated with true productivity, outside the NBA. More likely height confers an interpersonal advantage, so the return to height is a rent, at the expense of others. (Our culture has produced a song with the line: "Short people got no right to live.") Taxing height would indeed be odd (Greg Mankiw is pretty tall, by the way); but it is too nonchalant to presume that all market incomes reflect true productivity, let alone social productivity.

Fourth, evidence has accumulated that the degree of intergenerational income mobility in the United

States is less than it used to be, less than in some other advanced economies, and less than the American self-image requires. These changes and differences are unlikely to be genetic in origin. (If they were, they would be rents!) Mankiw ignores or skates over this evidence in favor of an anecdote: he thinks his children do not have significantly better opportunities than his own, although they grew up in a much more affluent family than he did. This is a rather small sample, of course. Besides, Professor Mankiw's success came in what must be one of the more meritocratic occupations around. I wonder if he would have had as good a shot at a bank presidency as his children could have.

Fifth, you don't have to be a card-carrying utilitarian to believe that taking a (lump-sum) dollar from a random rich person and giving it to a random poor person would lead to a better social state. Mankiw's attempts to undermine this intuition fall pretty far short: a) Yes, rich countries are pretty cheap with aid to poor countries, and might be so even if they were surer that aid was effective. But this only says something about the limits of human solidarity: greater within a family than outside it, greater within a nation than outside it, etc. So what? b) No, we don't try to equalize the marginal value of kidneys across the population. But this does not deny that some people need a kidney more than others, it merely reflects the fact that most people regard kidneys and other body parts as somehow less fungible than bank accounts. Does it follow from the fact that we do not redistribute spare kidneys that we should not redistribute some spare cash? As for the actual progressiveness of the federal tax code, if we take the group averages given in the paper as points on the tax function, the marginal tax rate between incomes of $223,500 and $1,219,700 appears to be slightly *lower* than that between $64,300 and $223,500.

Sixth, who could be against allowing people their "just deserts?" But there is that matter of what is "just." Most serious ethical thinkers distinguish between deservingness and happenstance. Deservingness has to be rigorously earned. You do not "deserve" that part of your income that comes from your parents' wealth or connections or, for that matter, their DNA. You may be born just plain gorgeous or smart or tall, and those characteristics add to the market value of your marginal product, but not to your just deserts. It may be impractical to separate effort from happenstance numerically, but that is no reason to confound them, especially when you are thinking about taxation and redistribution. That is why we may want to temper the wind to the shorn lamb, and let it blow on the sable coat.

Robert Solow
Massachusetts Institute of Technology
Cambridge, Massachusetts

References

Goldin, Claudia, and Lawrence F. Katz. 2008. *The Race between Education and Technology*. Belknap Press.

Mankiw, N. Gregory. 2013. "Defending the One Percent." *Journal of Economic Perspectives* 27(3): 21–34.

Murphy, Kevin J. 2012. "Pay, Politics, and the Financial Crisis." Chap. 12 in *Rethinking the Financial Crisis*, edited by Alan S. Blinder, Andrew W. Lo and Robert M. Solow. Russell Sage Foundation.

Response from N. Gregory Mankiw

Robert Solow's scattershot letter offers various gripes about my paper "Defending the One Percent." Let me respond, as blandly and cheerfully as I can, to his points.

First, Solow objects to my invoking the iconic Steve Jobs. Jobs is indeed an extreme case, but he also exemplifies much of what has happened to the US economy in recent years. In the same *JEP* issue as my paper, Kaplan and Rauh (2013) examine alternative hypotheses about increasing inequality. They conclude that "the increase in pay at the highest income levels is broad based" and that "the US evidence on income and wealth shares for the top 1 percent is most consistent with a 'superstar'-style explanation rooted in the importance of scale and skill-biased technological change." Solow says that I waffle on the role of finance, and about that, he is right. The social value of financial activity is hard to measure. When the evidence is inconclusive, honesty requires that we admit it.

Second, Solow is concerned about the political influence of the rich. I am less worried, in part because the wealthy include supporters of both the right (the Koch brothers) and the left (George Soros). Moreover, despite rising inequality, in 2008 and 2012 the United States managed to elect a left-leaning president committed to increasing taxes on the rich. In any event, the absence of such political concerns in my paper is hardly unique. The economics literature on optimal income redistribution, including the celebrated Mirrlees (1971) model, is based largely on utilitarian principles, not on some conjectured link between the income distribution and the electoral process.

Third, Solow thinks I have too narrow a conception of rents. Actually, I am less concerned about rents (incomes paid to factors of production in excess of opportunity cost) than I am about rent-seeking (the attempt to obtain rents by manipulating the political environment). If a person supplies his labor inelastically, his earnings are entirely rent, but that does not mean he is engaged in socially unproductive rent-seeking or that he is earning more than the value of his marginal product. Regarding the narrow, tangential question of why tall people earn higher wages, the literature has examined various hypotheses; for example, Case and Paxson (2008) report a positive correlation between height and cognitive skills.

Fourth, Solow interprets the evidence on intergenerational mobility as showing that the economy is not very meritocratic. (Oddly, he exempts the economics profession. He seems to believe that lack of success is often the result of bad luck or a rigged system, unless you are an economist, in which case it's your own fault.) Although I noted in my article that those born into extreme poverty face particularly difficult obstacles, I view the rest of the economy as more meritocratic than Solow does. In addition to the Kaplan and Rauh study, I recommend a popular book called *The Millionaire Next Door* (Stanley and Danko 1996). Written by two marketing professors who extensively surveyed high net worth individuals, the book reports that the typical millionaire is not someone who was born into wealth but rather is someone who has worked hard and lived frugally.

Moreover, the fact that higher income inequality is associated with lower intergenerational mobility is not a surprise. Consider the following assumptions: 1) A person's income Y depends on talent T and noise ε:

$$Y = \alpha T + \varepsilon.$$

2) Talent T is partly heritable, while noise ε is uncorrelated across generations. 3) Talent T and noise ε have constant variance but the return to talent α varies over time and across nations. Under these assumptions, higher income inequality goes hand in hand with lower intergenerational mobility. This simple statistical model does not explain why the return to talent varies, but it does explain why inequality and mobility often move in opposite directions.

Fifth, Solow tries to revive utilitarian logic without calling himself a utilitarian by claiming that redistribution leads to a better "social state." But what he means by social state, other than something like total utility, is unclear. In response to my critiques of utilitarianism, he says: a) people don't endorse more foreign aid because they don't feel solidarity with those in other nations, and b) people aren't in favor of kidney redistribution because they don't view kidneys as fungible. I agree. In fact, Solow's arguments seem more like restatements of my observations than refutations of them. If people were maximizing a conventional social welfare function behind a veil of ignorance, they would treat foreigners equally with their fellow citizens and they would treat kidneys as fungible. That they do not do so is evidence that our innate moral intuitions are far from utilitarian.

As for the progressivity of the actual tax system: Progressivity is correctly gauged by average tax rates, which rise strongly with income. Solow's reference to marginal tax rates is perplexing, as these are relevant for measuring incentive effects, not progressivity.

Sixth, and finally, Solow asks, who could be against allowing people their "just deserts"? Actually, much of the economics literature on redistribution takes precisely that stand, albeit without acknowledging doing so. The standard model assumes something like a utilitarian objective function and concludes that the optimal tax code comes from balancing diminishing marginal utility against the adverse incentive effects of redistribution. In this model, what people deserve plays no role in the formulation of optimal policy. I agree with Solow that figuring out what people deserve is hard, and I don't pretend to have the final word on the topic. But if my paper gets economists to focus a bit more on just deserts when thinking about policy, I will feel I have succeeded.

N. Gregory Mankiw
Harvard University
Cambridge, Massachusetts

References

Case, Anne, and Christina Paxson. 2008. "Stature and Status: Height, Ability, and Labor Market Outcomes." *Journal of Political Economy* 116(3): 499–532.

Kaplan, Steven N., and Joshua Rauh. 2013. "It's the Market: The Broad-Based Rise in the Return to Top Talent." *Journal of Economic Perspectives* 27(3): 35–56.

Mirrlees, James A. 1971. "An Exploration in the Theory of Optimal Income Taxation." *Review of Economic Studies* 38(2): 175–208.

Stanley, Thomas J., and William D. Danko. 1996. *The Millionaire Next Door: Surprising Secrets of America's Wealthy*. Marietta, GA: Longstreet Press.

***

It is conventional wisdom that income inequality grew almost unabated from the late 1970s up through the start of the Great Recession, with the top 1 percent controlling an increasing share of income in the United States and other English-speaking countries. The intellectual underpinnings of this now dominant view grew out of the tax-record-based research collected by Facundo Alvaredo, Anthony B. Atkinson, Thomas Piketty, and Emmanuel Saez in their world top incomes database and discussed in "The Top 1 Percent in International and Historical Perspective" (Summer 2013, pp. 3–20).

Tax records offer researchers many advantages, but their inherent disadvantages can skew inequality levels and affect their trends. Because the definitions of income are determined administratively, they vary over time within and across countries, and often differ from how most economists would define income. For example, while most economists agree that Social Security and other public transfers are income for individuals, tax record data largely miss this income in the United States and to a lesser degree in other countries. Similarly, individuals' business income is a real resource, whether reported on their personal income tax form, and thus captured in individual tax return data, or reported on a corporate income tax form and thus missed in the individual tax return data.

In Burkhauser et al. (2012), we raise these measurement concerns as they relate to the unusually large increase in inequality in personal tax data between 1986 and 1988 following the 1986 Tax Reform Act. In Piketty and Saez's top income series excluding capital gains (the series they focus on in Piketty and Saez 2003), the increase in the top 1 percent's income shares during the two years following the implementation of the act represents approximately one-third of the total increase since 1985. This change appears to reflect a reclassification from corporate to individual income in response to top individual tax rates falling below the corporate rate, rather than a true inequality change. Similarly, Burkhauser, Hahn, and Wilkins (2013) note that a short-term spike in Australian top incomes in tax-record data between 1986 and 1989 appears to be driven by a change in the treatment of company profits and dividends that more fully captured company profits in the personal income tax base. The longer-term growth in inequality in tax record data since 1989 seems to arise because Australia started taxing longer-term realized capital gains in 1989—but only on property purchased after September 19, 1985.

Alvaredo, Atkinson, Piketty, and Saez acknowledge that tax changes may affect income measures over time. But in the US case, they dismiss this concern by referencing their data series that includes realized taxable capital gains. They claim doing so eliminates a major part of the tax avoidance channel, because corporate income still appears as capital gains. This rebuttal would be a stronger if realized taxable capital gains had a close relationship to current income. They do not. Because the United States only taxes capital gains when the asset is sold, individuals can defer realizing them indefinitely. Thus, capital gains may not appear on individual tax forms until years or decades after being generated, if at all. This also means corporate income before the 1986 Act may appear as realized capital gains in the 1990s or 2000s. Hence, including taxable realized capital gains as a measure of current market returns to capital can importantly affect such trends. The classic Haig–Simons income definition includes capital income, but at the point of accrual, not when such gains are realized.

In Burkhauser et al. (2012), with data from the Current Population Survey, we largely replicate the inequality trends since the 1960s found by Piketty and Saez (2003) for market income. But when we extend our analysis and use a more comprehensive post-tax, post-transfer income measure, including in-kind income and accrued capital gains, we observe different income trends (Armour, Burkhauser, and Larrimore 2013). Current levels of income inequality are still at or near record levels. But inequality has not appreciably increased since the late 1980s or early 1990s.

This finding does not discount Alvaredo and colleagues' insights on the causes of increased market income inequality. In many ways, their explanations still apply when using accrued, rather than realized, capital gains. If increased bargaining by executives and capital owners due to lower tax rates is important, then market income inequality should have increased more in the 1980s, when top income tax rates fell, than in the 1990s, when top income tax rates increased. This pattern is likely to be the case when using accrued capital gains, but is less true when using realized capital gains—which partially shift the timing of these returns to capital into the 1990s and 2000s.

But designing policy recommendations based on pre-tax, pre-transfer market income trends over long periods of time seems misguided. If you compare distributions of market income in the 1920s when taxes and government transfers were low with market income today, you miss the dramatic growth of progressive income taxes and transfers like Social Security and the Earned Income Tax Credit that shifted income to people with less market income in the real world. More subtly, leaving out the real world behavioral consequences of the growth of Social Security and other government transfers obfuscates the impact these policies have on market income. The US labor force participation rate of men aged 65 and over was well over 50 percent in the 1920s and is now under 20 percent. In a market income world, the elderly look much worse off today, but in fact their income has shifted to nonmarket sources. In this sense, social policies may increase measured market income inequality even if they reduce inequality under broader income measures. Thus, while research using tax records contributes greatly to the inequality debate, it should be viewed as an addition to, rather than a replacement for, the extensive research conducted on both US and international inequality trends using broader income measures.

Richard V. Burkhauser
Cornell University
Ithaca, New York

Jeff Larrimore
Joint Committee on Taxation
Washington, DC

References

Armour, Philip, Richard V. Burkhauser, and Jeff Larrimore. 2013. "Levels and Trends in United States Income and Its Distribution: A Crosswalk from Market Income Towards a Comprehensive Haig–Simons Income Approach." NBER Working Paper 19110.

Burkhauser, Richard V., Shuaizhang Feng, Stephen P. Jenkins, and Jeff Larrimore. 2012. "Recent Trends in Top Income Shares in the USA: Reconciling Estimates from March CPS and IRS Tax Return Data." *Review of Economics and Statistics* 94(2): 371–88.

Burkhauser, Richard V., Markus H. Hahn, and Roger Wilkins. 2013. "Measuring Top Incomes Using Tax Record Data: A Cautionary Tale from Australia." NBER Working Paper 19121.

Facundo Alvaredo, Anthony B. Atkinson, Thomas Piketty, and Emmanuel Saez. 2013. "The Top 1 Percent in International and Historical Perspective." *Journal of Economic Perspectives* 27(3): 3–20.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118(1): 1–39.

Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 249–250

Notes

For additional announcements, check out the continuously updated JEP online Bulletin Board (http://www.aeaweb.org/bulletinboard.php). Calls for papers, notices of professional meetings, and other announcements of interest to economists should be submitted to Ann Norman, at jep@jepjournal.org, in one or two paragraphs containing the relevant information. These will be posted at the JEP online Bulletin Board. Given sufficient lead time (at least one month before an issue goes online), we will also print a shorter, one-paragraph version of your notice in the "Notes" section of the Journal of Economic Perspectives. *We reserve the right to edit material received.*

Call for Sessions and Papers for the 2015 AEA Annual Meeting. The next meeting will be held January 3–5, 2015, in Boston, MA. Members wishing to give papers or organize complete sessions for the program in Boston are invited to submit proposals electronically to Professor Richard H. Thaler via the American Economic Association web site http://www.aeaweb.org/Annual_Meeting/, beginning March 1st.

To be considered, **individual paper proposals should be submitted by April 1, 2014**, (with abstracts) and up to *two Journal of Economic Literature bibliographic codes* in rank order. The **deadline for complete session proposals is April 15, 2014**. *At least one author of each paper must be an AEA member.* All authors of papers on a complete session must join the AEA if the session is selected for the program.

Proposals for complete sessions have historically had a higher probability of inclusion (35–40%) than papers submitted individually (10–15%). Individual paper contributors are strongly encouraged to use the AEA's Econ-Harmony website (http://www.aeaweb.org/econ-harmony/) to form integrated sessions. Proposals for a complete session should be submitted only by the session organizer. Sessions normally contain three or four papers.

Please make certain your information is complete before submission. No changes will be accepted until a decision is made about inclusion on the program (usually in July). Econometric studies or highly mathematical papers are not appropriate for sessions sponsored by the AEA; such papers should be submitted to the Econometric Society. Do not send a complete paper. The Association discourages multiple proposals from the same person, and under no circumstances should the same person submit more than two proposals.

Some of the papers presented at the annual meeting are published in the May *American Economic Review* (the *Papers & Proceedings*). The President-elect includes at least three contributed sessions (12 papers) from among those submitted in response to this call for sessions and papers.

Econ-Harmony helps prospective individual paper submitters for the AEA Annual Meetings find others with similar interests who might join them to form a complete session submission, and provides an opportunity to volunteer as a session chair. Twenty-five percent of submitted complete sessions and 14% of submitted individual papers appeared on the 2014 Program; 31% of submitted complete sessions and 12% of submitted individual papers appeared on the 2013 Program. Econ-Harmony is available at http://www.aeaweb.org/econ-harmony/.

Webcasts are Online. Eleven 2014 Annual Meeting sessions are available online:

AEA Presidential Address: "A Grand Gender Convergence: Its Last Chapter" (Claudia Goldin)
AEA Awards Ceremony (William Nordhaus)
Richard T. Ely Lecture: "Retirement Security in an Aging Population" (James Poterba)
Nobel Laureate Luncheon (Paul Milgrom, Roger Myerson, and Alvin Roth)
AEA/AFA Joint Luncheon (Jeremy Stein)
Chairman Bernanke Presentation (Ben Bernanke, Kenneth Rogoff, and Anil Kashyap)
"What's Natural? Key Macroeconomic Parameters after the Great Recession"
"Discounting for the Long Run"
"Financial Globalization"
"Climate Change Policy after Kyoto"
"Macroeconomics of Austerity"

Go to http://www.aeaweb.org/webcasts/2014/index.php. Webcasts from 2009 through 2014 are available to all compliments of the AEA. Webcasts for the Continuing Education Programs, including 2014, are available for members only. See the link at the bottom of the webpage.

Congratulations to Janet Yellen who is now confirmed as the next Federal Reserve Chair! Ms. Yellen is a 2012 AEA Distinguished Fellow and former Vice President of the Association. Her long-standing influence on the economics profession is demonstrated by her participation in various AEA activities including research published in AEA journals. Five articles authored by Janet Yellen are available open-access at http://www.aeaweb.org/yellen.php.

Nominations for *American Economics Journal* Best Article Prizes. We solicit nominations for best paper prizes for each of the *American Economic Journals*. Awards will be announced in April 2014, selected from articles published in the 2011–2013 volumes of each *AEJ*. Suggest papers **by February 15, 2014,** at http://www.aeaweb.org/aej/bestpapernominations.php.

The Fourth Annual AEA Conference on Teaching and Research in Economic Education (CTREE) will be held May 28–30, 2014, at the Washington Marriott at Metro Center. The conference is hosted by the AEA Committee on Economic Education in cooperation with the *Journal of Economic Education*. Plenary talks will be given by Alan Blinder (Princeton), Kenneth G. Elzinga (UVA), Cecilia Rouse (Princeton), and others to be announced. Conference registration will open February 15, 2014. For more information go to http://www.aeaweb.org/committees/AEACEE/index.php.

Call for Artists. The program cover for the ASSA annual meeting often features artwork by economists or their spouses. Please contact us if you would like to see your artwork on a cover or have suggestions for future covers. Artwork can be provided in any medium that can be photographed. If selected, a color transparency, slide, or photograph (color or black and white) will be required. There is no remuneration, just the honor of having your work recognized. Send suggestions by August 15, 2014, to: gwyn.p.loftis@vanderbilt.edu or ASSA Cover, American Economic Association, 2014 Broadway, Suite 305, Nashville, TN, 37203.

Stay in Touch with CSWEP. The *CSWEP News* keeps you up to date on the progress of women in the economics profession and on professional opportunities such as the Bell Award, the Bennett Prize, and upcoming mentoring workshops. *CSWEP News* also includes calls for papers at national and regional meetings, for applications to the Summer Fellows Program, and for registration to the Annual Mentoring Breakfasts at ASSA. Much of this information is of particular interest to younger economists, regardless of gender. Digital subscriptions to the *CSWEP News* are free and automatically accord the status of "CSWEP Affiliate" to subscribers. Find out more at http://www.aeaweb.org/committees/cswep/index.php.

Retired Faculty Available for Part-time or Temporary Teaching. JOE now lists retired economists interested in teaching on either a part-time or temporary basis at http://www.aeaweb.org/joe/available_faculty/. Individuals can add or delete their name at any time. Listings are deleted on November 30. The service is closed during December and January, and re-opens on February 1.

Publication Updates on Twitter. Want to see forthcoming article previews as soon as they are available for the *American Economic Review*, *American Economic Journals*, and the *Journal of Economic Literature*? Follow us on Twitter! At https://twitter.com/AEAjournals/.

2014 Application/Renewal for Membership

AMERICAN ECONOMIC ASSOCIATION
2014 Broadway, Suite 305
Nashville, TN 37203
Ph. 615-322-2595 fax: 615-343-7590
Federal ID No. 36-2166945
www.vanderbilt.edu/AEA

RENEWING MEMBERS, ENTER ACCT. NUMBER & EXP. DATE	IF PAYING BY CREDIT CARD, PLEASE FILL OUT BELOW
ACCOUNT NUMBER:	CARD NUMBER:
EXPIRATION DATE:	EXP DATE: CSC CODE:

FIRST NAME:	MI:	LAST NAME:
ADDRESS:		
CITY:		STATE/PROVINCE: ZIP:
COUNTRY:		☐ Check here if non-US
PHONE:		FAX:
PRIMARY FIELD OF SPECIALIZATION:		
SECONDARY FIELD OF SPECIALIZATION:		
EMAIL:		☐ Check here to exclude your email address from the public directory

Please include my email address to receive:
☐ Announcements about public policy affecting economists or the economics profession
☐ Surveys of economists for research purposes

MEMBERSHIP DUES — Based on annual income. Please select one below.

☐ Annual income of $70,000 or less	$20	$
☐ Annual income of $70,000 to $105,000	$30	$
☐ Annual income over $105,000	$40	$

The AEA dues above include online access to all seven AEA journals.
For print or CD subscription(s) indicate preference below and add appropriate charge(s).

Journal		Print	Int'l Postage*	CD*	
AER	(12 issues, incl. P&P)	☐ Add $25	☐ Add $35	☐ Add $25	$
AER Papers & Proceedings Only*		☐ Add $10	n/a	n/a	$
JEL	(4 quarterly issues)	☐ Add $15	☐ Add $15	☐ Add $15	$
JEP	(4 quarterly issues)	☐ Add $15	☐ Add $15	☐ Add $15	$
AEJ: Applied	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$
AEJ: Policy	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$
AEJ: Macro	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$
AEJ: Micro	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$

* Int'l postage applies only to print journals mailed outside of the U.S. No additional postage is required for CDs or the AER Papers and Proceedings.

AEA Journals via JSTOR online

JSTOR	☐ Add $16	$

				Sub Total	$
Check One:	☐ 1 Year	☐ 2 Years	☐ 3 Years	TOTAL AMOUNT	$

Make checks payable to: American Economic Association.
Must be drawn on a US bank.
Apply online at http://www.aeaweb.org/membership.php

Payments must be made in advance. We accept checks (in US dollars only, with correct coding for processing in US banks) and credit cards; online or by faxing or mailing the application. Please choose one method; it is the Association's policy NOT TO REFUND dues.

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.

FSC
www.fsc.org
MIX
Paper from responsible sources
FSC® C101537

The Journal of

Economic Perspectives

Winter 2014, Volume 28, Number 1

AMERICAN ECONOMIC ASSOCIATION