The Journal of Economic Perspectives

A journal of the American Economic Association

Winter 2015

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 2015, Vol. 29, 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 29 • Number 1 • Winter 2015

Symposium

Wealth and Inequality

Articles

Features

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 3–28

The Rise and Decline of General Laws of Capitalism[†]

Daron Acemoglu and James A. Robinson

Economists have long been drawn to the ambitious quest of discovering the general laws of capitalism. David Ricardo, for example, predicted that capital accumulation would terminate in economic stagnation and inequality as a greater and greater share of national income accrued to landowners. Karl Marx followed him by forecasting the inevitable immiseration of the proletariat. Thomas Piketty's (2014) tome, *Capital in the Twenty-First Century*, emulates Marx in his title, his style of exposition, and his critique of the capitalist system. Piketty is after general laws that will demystify our modern economy and elucidate the inherent problems of the system—and point to solutions.

But the quest for general laws of capitalism is misguided because it ignores the key forces shaping how an economy functions: the endogenous evolution of technology and of the institutions and the political equilibrium that influence not only technology but also how markets function and how the gains from various different economic arrangements are distributed. Despite his erudition, ambition, and creativity, Marx was led astray because of his disregard of these forces. The same is true of Piketty's sweeping account of inequality in capitalist economies.

In the next section, we review Marx's conceptualization of capitalism and some of his general laws. We then turn to Piketty's approach to capitalism and his general laws. We will point to various problems in Piketty's interpretation of the economic relationships underpinning inequality, but the most important shortcoming is that,

■ *Daron Acemoglu is Elizabeth and James Killian Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. James A. Robinson is Wilbur A. Cowett Professor of Government, Harvard University, Cambridge, Massachusetts. Their email addresses are daron@mit.edu and jrobinson@gov.harvard.edu.*

[†] To access the Appendix and Data Appendix, visit
http://dx.doi.org/10.1257/jep.29.1.3

doi=10.1257/jep.29.1.3

though he discusses the role of certain institutions and policies, he allows neither for a systematic role of institutions and political factors in the formation of inequality nor for the endogenous evolution of these institutional factors. This implies that his general laws have little explanatory power. We illustrate this by first using regression evidence to show that Piketty's central economic force, the relationship between the interest rate and the rate of economic growth, is not correlated with inequality (in particular, with a key variable he focuses on, the share of national income accruing to the richest 1 percent, henceforth, the top 1 percent share). We then use the examples of the South African and Swedish paths of inequality over the 20th century to demonstrate two things: First, that using the top 1 percent share may miss the big picture about inequality. Second, it is impossible to understand the dynamics of inequality in these societies without systematically bringing in institutions and politics and their endogenous evolution. We conclude by outlining an alternative approach to inequality that eschews general laws in favor of a conceptualization in which both technology and factor prices are shaped by the evolution of institutions and political equilibria—and institutions themselves are endogenous and are partly influenced by, among other things, the extent of inequality. We then apply this framework to the evolution of inequality and institutions in South Africa and Sweden.

We should note at this point that we do not believe the term capitalism to be a useful one for the purposes of comparative economic or political analysis. By focusing on the ownership and accumulation of capital, this term distracts from the characteristics of societies which are more important in determining their economic development and the extent of inequality. For example, both Uzbekistan and modern Switzerland have private ownership of capital, but these societies have little in common in terms of prosperity and inequality because the nature of their economic and political institutions differs so sharply. In fact, Uzbekistan's capitalist economy has more in common with avowedly noncapitalist North Korea than Switzerland, as we argued in Acemoglu and Robinson (2012). That said, given the emphasis in both Marx and Piketty on capitalism, we have opted to bear with this terminology.

Capital Failures

Though many important ideas in social science can be traced to Karl Marx's oeuvre, his defining approach was to seek certain hard-wired features of capitalism—what Marx called general laws of capitalist accumulation. This approach was heavily shaped by the historical context of the middle 19th century in which Marx lived and wrote. Marx experienced first-hand both the bewildering transformation of society with the rise of industrial production, and the associated huge social dislocations.

Marx developed a rich and nuanced theory of history. But the centerpiece of this theory, historical materialism, rested on how material aspects of economic life, together with what Marx called forces of production—particularly technology—shaped all other aspects of social, economic, and political life, including the relations of production. For example, Marx famously argued in his 1847 book *The Poverty of*

Philosophy that “the hand-mill gives you society with the feudal lord; the steam-mill society with the industrial capitalist” (as reprinted in McLellan 2000, pp. 219–220). Here the hand-mill represents the forces of production while feudalism represents the relations of production, as well as a specific set of social and political arrangements. When the forces of production (technology) changed, this destabilized the relations of production and led to contradictions and social and institutional changes that were often revolutionary in nature. As Marx put it in 1859 in *A Contribution to the Critique of Political Economy* (McLellan 2000, p. 425):

> [T]he sum total of these relations of production constitutes the economic structure of society—the real foundation, on which rise legal and political superstructures and to which correspond definite forms of social consciousness. The mode of production of material life conditions the general character of the social, political and spiritual processes of life. At a certain state of their development the material forces of production in society come into conflict with the existing relations of production or—what is but a legal expression of the same thing—with the property relations within which they had been at work before. From forms of development of the forces of production these relations turn into fetters. Then comes the epoch of social revolution. With the change of the economic foundation the entire immense superstructure is more or less rapidly transformed.

Marx hypothesized that the forces of production, sometimes in conjunction with the ownership of the means of production, determined all other aspects of economic and political institutions: the de jure and de facto laws, regulations, and arrangements shaping social life. Armed with this theory of history, Marx made bold predictions about the dynamics of capitalism based just on economic fundamentals—without any reference to institutions or politics, which he generally viewed as derivative of the powerful impulses unleashed by the forces of production.[1]

Most relevant for our focus are three of these predictions concerning inequality. In *Capital* (1867, Vol. 1, Chap. 25), Marx developed the idea that the reserve army of the unemployed would keep wages at subsistence level, making capitalism inconsistent with steady improvements in the living standards of workers. His exact prediction here is open to different interpretations. Though Marx (1867, Vol. 1,

[1] There is no consensus on Marx’s exact formulation of the relationship between the “substructure,” comprising productive forces and sometimes the relations of production, and the “superstructure” which includes what we call political institutions and most aspects of economic institutions. In Chapter I of the *Communist Manifesto*, Marx and Engels wrote that “The history of all hitherto existing society is the history of class struggles.” But the idea here, so far as we understand, is not that “class struggle” represents some autonomous historical dynamic, but rather that it is an outcome of the contradictions between the forces of production and the ownership of the means of production. In some writings, such as *The Eighteenth Brumaire of Louis Napoleon*, Marx also allowed for feedback from politics and other aspects of society to the forces of production. But it is clear from his work that he regraded this as second order (see Singer 2000, chapter 7 for a discussion of this). Marx never formulated an approach in which institutions play the central role and themselves change endogenously.

Chapter 25, Section 3) viewed capitalism as the harbinger of "misery, agony of toil, slavery, ignorance, brutality, and mental degradation" for working men, it is less clear whether this was meant to rule out real wage growth. Blaug (1996) states that Marx never claimed that real wages would be stagnant, but rather that the share of labor in national income would fall since Marx (1867, Vol 1., Chapter 24, Section 4) says "real wages . . . never rise proportionately to the productive power of labor." Foley (2008, Chapter 3), on the other hand, argues that Marx did start by asserting that real wages would not rise under capitalism, but then weakened this claim to a falling labor share when he realized that wages were indeed increasing in Britain. This motivates us to state this law in both a strong and a weak form. Under either its strong or weak form, this law implies that any economic growth under capitalism would almost automatically translate into greater inequality—as capitalists benefit and workers fail to do so. We combine this with a second general law of capitalism from Volume III of *Capital* and a third law, less often stressed but highly relevant, presented in Volume I of *Capital*. Thus, three key predictions from Marx are:

1) *The General Law of Capitalist Accumulation.* Strong Form**:** Real wages are stagnant under capitalism. Weak Form: The share of national income accruing to labor would fall under capitalism.

2*) The General Law of Declining Profit*: as capital accumulates, the rate of profit falls.

3) *The General Law of Decreasing Competition*: capital accumulation leads to increased industrial concentration.

Marx's general laws did not fare well. As Marx was writing, real wages, which had been constant or falling during the first decades of the 19th century, had already been rising, probably for about two decades (Allen 2001, 2007, 2009a; Clark 2005; Feinstein 1998). The share of labor in national income, which had fallen to under half by 1870, also started to increase thereafter, reaching two-thirds in the 20th century. Allen's (2009a) calculation of the real rate of profit suggests that the profit rate was comparatively low at the end of the 18th century and rose until around 1870 reaching a maximum of 25 percent, but then fell back to around 20 percent, where it stabilized until World War I. Matthews, Feinstein, and Odling-Smee (1982, pp. 187–88) suggest that these rates did not fall in the 20th century, though there is a lot of heterogeneity across sectors. (The third law's performance was no better as we discuss below.)

Why did Marx's general laws fail? Mostly because they ignored both the endogenous evolution of technology (despite his great emphasis on the forces of production) and also the role of institutions and politics that shape markets, prices, and the path of technology. The increase in real wages in Britain, for example, was in part a consequence of the change in the pace and nature of technological change, rapidly increasing the demand for labor (Crafts 1985; Allen 2009b; Mokyr 2012).

The rationalization of property rights, dismantling of monopolies, investment in infrastructure, and the creation of a legal framework for industrial development, including the patent system, were among the institutional changes contributing to rapid technological change and its widespread adoption in the British economy (Acemoglu and Robinson 2012; Mokyr 2012).

The distribution of the gains from new technologies was also shaped by an evolving institutional equilibrium. The Industrial Revolution went hand-in-hand with major political changes, including the development of the state and the Reform Acts of 1832, 1867, and 1884, which transformed British political institutions and the distribution of political power. For example, in 1833 a professional factory inspectorate was set up, enabling the enforcement of legislation on factory employment. The political fallout of the 1832 democratization also led in 1846 to the repeal of the Corn Laws (tariffs limiting imports of lower-priced foreign corn), lowering the price of bread, raising real wages, and simultaneously undermining land rents (Schonhart-Bailey 2006). The Factory Act of 1847 took the radical step of limiting working hours in the textile mills to ten hours per day for women and teenagers. The Reform Act of 1867 led to the abolition of the Masters and Servants Acts in 1875—which had imposed on workers legally enforceable duties of loyalty and obedience, and limited mobility—illustrating the role of pro-worker labor market legislation that increased real wages (Naidu and Yuchtman 2013).

Another telling example is the failure of Marx's third general law in the United States: the prediction of increased industrial concentration. After the end of the US Civil War came the age of the robber barons and the huge concentration of economic ownership and control. By the end of the 1890s, companies such as Du Pont, Eastman Kodak, Standard Oil, and International Harvester came to dominate the economy, in several cases capturing more than 70 percent of their respective markets (Lamoreaux 1985, pp. 3–4). It looked like a Marxian prediction come true—except that this situation was transitory and was duly reversed as popular mobilization, in part triggered by the increase in inequality, changed the political equilibrium and the regulation of industry (Sanders 1999). The power of large corporations started being curtailed with the Interstate Commerce Act of 1887 and then the Sherman Anti-Trust Act of 1890, which were used in the early 20th-century trust-busting efforts against Du Pont, the American Tobacco Company, the Standard Oil Company, and the Northern Securities Company, then controlled by J.P. Morgan. The reforms continued with the completion of the break-up of Standard Oil in 1911; the ratification of the Sixteenth Amendment in 1913, which introduced the income tax; and the Clayton Anti-Trust Act in 1914 and the founding of the Federal Trade Commission. These changes not only stopped further industrial concentration but reversed it (Collins and Preston 1961; Edwards 1975). White (1981) shows that US industrial concentration in the post–World War II period changed little (see White 2002 for an update).

Crucially, the political process that led to the institutional changes transforming the British economy and inequality in the 19th century was not a forgone conclusion. Nor was the rise in inequality in 19th century United States after its Civil War an inevitable consequence of capitalism. Its reversal starting in the early 1900s was equally dependent on an evolving institutional equilibrium. In fact, while the power of monopoly and inequality were being curtailed in the United States, inequality continued to increase rapidly in neighboring Mexico under the authoritarian rule of Porfirio Diaz, culminating in revolution and civil war in 1910, and demonstrating the central role of the endogenous and path-dependent institutional dynamics.

Marx's general laws failed for the same reason that previous general laws by other economists also performed poorly. These laws were formulated in an effort to compress the facts and events of their times into a grand theory aiming to be applicable at all times and places, with little reference to institutions and the (partly institutionally determined) changing nature of technology. For example, when David Ricardo published the first edition of *On the Principles of Political Economy and Taxation* in 1817, and predicted that a rising share of national income would accrue to land, he had indeed been living through a period of rapidly rising land rents in Britain. But soon thereafter, the share of national income accruing to land started a monotonic decline, and by the 1870s real rents started a rapid fall, which would last for the next 60 years (Turner, Beckett, and Afton 1999; Clark 2002, 2010).

In short, Marx's general laws, like those before him, failed because they relied on a conception of the economy that did not recognize the endogenous evolution of technology and the role of changing economic and political institutions, shaping both technology and factor prices. In fact, even Marx's emphasis on the defining role of the forces of production, so emblematic of his approach, was often inadequate not only as the engine of history, but also as a description of history, including his paradigmatic example of hand-mills and steam-mils. For example, Bloch (1967) argued persuasively that the hand-mill did not determine the nature of feudal society, nor did the steam-mill determine the character of the post-feudal world.

Seeking 21st-Century Laws of Capitalism

Thomas Piketty is also an economist of his milieu, with his thinking heavily colored by increasing inequality in the Anglo-Saxon world and more recently in continental Europe—and in particular compared to the more equal distribution of labor and total incomes seen in France in the 1980s and 1990s. A large literature in labor economics had done much to document and dissect the increase in inequality that started sometime in the 1970s in the United States (see the surveys and the extensive references to earlier work in Katz and Autor 1999 and Acemoglu and Autor 2011). This literature has demonstrated that the increase in inequality has taken place throughout the income distribution and that it can be explained reasonably well by changes in the supply and demand for skills and in labor market institutions. Piketty and Saez (2003) brought a new and fruitful perspective to this literature by using data from tax returns, confirming and extending the patterns the previous literature had uncovered and placing a heavy emphasis on rising inequality at the very top of the income distribution.

In *Capital in the Twenty-first Century*, Piketty goes beyond this empirical and historical approach to offer a theory of the long-run tendencies of capitalism. Though Piketty's data confirm the finding of the previous literature that widening inequality in recent decades, at least in advanced economies, had been driven by rising inequality of labor incomes, his book paints a future dominated by capital income, inherited wealth, and rentier billionaires. The theoretical framework used to reach this conclusion is a mix of Marxian economics with Solow's growth model. Piketty defines capitalism in the same way that Marx does, and has a similarly materialist approach linking the dynamics of capitalism to the ownership of the means of production (in particular capital) and the ironclad nature of technology and exogenous growth dynamics. It is true that Piketty sometimes mentions policies and institutions (for example, the wealth tax and the military and political developments that destroyed capital and reduced the ratio of wealth to income during the first half of the 20th century). But their role is ad hoc. Our argument is that, to explain inequality, these features and their endogenous evolution have to be systematically introduced into the analysis.

This approach shapes Piketty's analysis and predictions about the nature of capitalism. *Capital in the Twenty-first Century* starts by introducing two "fundamental laws," but the more major predictions flow from what Piketty calls a "fundamental force of divergence" (p 351) or sometimes the "fundamental inequality" (p. 25), comparing the (real) interest rate of the economy to the growth rate.

The first fundamental law is just a definition:

$$\text{capital share of national income} = r \times (K/Y),$$

where r is the net real rate of return on capital (which can be viewed as a real interest rate), K is the capital stock, and Y is GDP (or equivalently, national income as the economy is taken to be closed).

The second fundamental law is slightly more substantial. It states that

$$K/Y = s/g,$$

where s is the saving rate and g is the growth rate of GDP. As we explain in the online Appendix (available with this paper at http://e-jep.org), a version of this law does indeed follow readily from the steady state of a Solow-type model of economic growth (but see Krusell and Smith 2014; Ray 2014). At an intuitive level, the growth rate of the capital stock K will be given by net investment, which in a closed economy will be equal to saving, sY. Thus, the ratio K/Y will reflect the ratio "change in K to change in Y" over time due to economic growth, which is s/g.

Let us follow Piketty here and combine these two fundamental laws to obtain

$$\text{capital share of national income} = r \times (s/g).$$

Piketty posits that, even as g changes, r and s can be taken to be approximate constants (or at least that they will not change as much as g). This then leads to what can be thought of as his first general law, that when growth is lower, the capital share of national income will be higher.

This first law is not as compelling as one might at first think, however. After all, one must consider whether a change in the growth rate g might also alter the saving rate s or the rate of return r, because these are all endogenous variables that are linked in standard models of economic growth. Piketty argues that r should not change much in response to a decline in g because the elasticity of substitution between capital and labor is high, resulting in an increase in the capital share of national income.[2]

However, the vast majority of existing estimates indicate a short-run elasticity of substitution significantly less than one (for example, Hamermesh 1993; Mairesse, Hall, and Mulkay 1999; Chirinko, Fazzari, and Meyer 1999; Krusell, Ohanian, Rios-Rull, and Violante 2000; Chirinko 1993; Antràs 2004; Klump, McAdam, and Willman 2007; Oberfield and Raval 2014). This is also the plausible case on intuitive grounds: given technology, the ability to substitute capital for labor would be limited (for example, if you reduce labor to zero, for a given production process, one would expect output to fall to zero also). Though this elasticity could be higher in longer horizons, Chirinko (2008) and Chirinko and Mallick (2014) find it to be significantly less than one also in the long run. One reason why the long-run elasticity of substitution might be greater than one is the endogeneity of technology (for example, Acemoglu 2002, 2003). In this context, it is worth noting that the only recent paper estimating an elasticity of substitution greater than one, Karabarbounis and Neiman (2014), uses long-run cross-country variation related to changes in investment prices, making their estimates much more likely to correspond to endogenous-technology elasticities. Nevertheless, as Rognlie (2014) points out, even an elasticity of substitution significantly greater than one would not be sufficient to yield the conclusions that Piketty reaches.

Moreover, though it is true that there has been a rise in the capital share of national income, this does not seem to be related to the forces emphasized in *Capital in the Twenty-First Century*. In particular, Bonnet, Bono, Chapelle, and Wasmer (2014) demonstrate that this rise in the capital share is due to housing and the increased price of real estate, shedding doubt on the mechanism Piketty emphasizes.

The second general law of *Capital in the Twenty-First Century* is formulated as

$$r > g,$$

stating that the (real) interest rate exceeds the growth rate of the economy. Theoretically, in an economy with an exogenous saving rate, or with overlapping generations (for example, Samuelson 1958; Diamond 1965), or with incomplete markets

[2] However, the interest rate and the growth rate are linked from both the household side and the production side. For example, with a representative household, we have that $r = \theta g + \rho$, where θ is the inverse of the intertemporal elasticity of substitution and ρ is the discount rate. The fact that the representative household assumption may not be a good approximation to reality does not imply that r is independent of g. On the production side, g affects r through its impact on the capital stock, and it is the second channel that depends on the elasticity of substitution between capital and labor.

(for example, Bewley 1986; Aiyagari 1994), the interest rate need not exceed the growth rate. It will do so in an economy that is *dynamically efficient*, meaning in an economy in which it is impossible to increase the consumption at all dates (thus achieving a Pareto improvement). Whether an economy is dynamically efficient is an empirical matter—for example, Geerolf (2013) suggests that several OECD economies might be dynamically inefficient—and dynamic inefficiency becomes more likely when the capital-output ratio is very high as *Capital in the Twenty-first Century* predicts it to be in the future.

Finally, Piketty's third and most important general law is that whenever $r > g$, there will be a tendency for inequality to rise. This is because capital income will tend to increase at the rate of interest, r, while national income (and the income of noncapitalists) increases at the rate g. Because capital income is unequally distributed, this will translate into a capital-driven increase in inequality, taking us back to the age of Jane Austen and Honoré Balzac. In the words of Piketty (pp. 25–26): "This fundamental inequality [$r > g$] will play a crucial role in this book. In a sense, it sums up the overall logic of my conclusions. When the rate of return on capital significantly exceeds the growth rate of the economy, then it logically follows that inherited wealth grows faster than output and income."

He elaborates on this point later, writing: "The primary reason for the hyperconcentration of wealth in traditional agrarian societies and to a large extent in all societies prior to World War I is that these were low-growth societies in which [sic] the rate of return on capital was markedly and durably higher than the rate of growth" (p. 351). Based on this, he proposes an explanation for the rise in inequality over the next several decades: "The reason why wealth today is not as unequally distributed as in the past is simply that not enough time has passed since 1945" (p. 372).[3]

As with the first two general laws, there are things to quibble with in the pure economics of the third general law. First, as already mentioned, the emphasis on $r - g$ sits somewhat uneasily with the central role that labor income has played in the rise in inequality. Second, as we show in the online Appendix, $r > g$ is fully consistent with constant or even declining inequality. Third, $r - g$ cannot be taken as a primitive on which to make future forecasts, as both the interest rate and the growth rate will adjust to changes in policy, technology, and the capital stock. Finally, in the presence of a modest amount of social mobility, even very large values of $r - g$ do not lead to divergence at the top of the distribution (again, as we show in the online Appendix).

But our major argument is about what the emphasis on $r > g$ leaves out: institutions and politics. Piketty largely dismisses the importance of institutions against the

[3] It is unclear whether $r > g$ is a force towards divergence of incomes across the distribution of income, or towards convergence to a new and more unequal distribution of income. In many places, including those we have already quoted, Piketty talks of divergence. But elsewhere, the prediction is formulated differently, for example, when he writes: "With the aid of a fairly simple mathematical model, one can show that for a given structure of . . . [economic and demographic shocks]. . ., the distribution of wealth tends towards a long-run equilibrium and that the equilibrium level of inequality is an increasing function of the gap $r - g$ between the rate of return on capital and the growth rate" (p. 364). In the online Appendix, we discuss a variety of economic models linking $r - g$ to inequality.

crushing force of the fundamental inequality, writing that "the fundamental inequality $r > g$ can explain the very high level of capital inequality observed in the 19th century, and thus in a sense the failure of the French Revolution. The formal nature of the regime was of little moment compared with the inequality $r > g$" (p. 365). In passing, we should note that the available empirical evidence suggests that the French Revolution not only led to a decrease in inequality (Morrisson and Snyder 2000), but also profoundly changed the path of institutional equilibria and economic growth in Europe (Acemoglu, Cantoni, Johnson, and Robinson 2011).

If the history of grand pronouncements of the general laws of capitalism repeats itself—perhaps first as tragedy and then farce as Marx colorfully put it—then we may expect the same sort of frustration with Piketty's sweeping predictions as they fail to come true, in the same way that those of Ricardo and Marx similarly failed in the past. We next provide evidence suggesting that this is in fact quite likely as the existing evidence goes against these predictions.

Cross-Country Data on $r > g$ and Top-Level Inequality

The major contribution of Piketty, often together with Emmanuel Saez, has been to bring to the table a huge amount of new data on inequality (Piketty and Saez 2003). The reader may come away from these data presented at length in Piketty's book with the impression that the evidence supporting his proposed laws of capitalism is overwhelming. However, Piketty does not present even basic correlations between $r - g$ and changes in inequality, much less any explicit evidence of a causal effect. Therefore, as a first step we show that the data provide little support for the general laws of capitalism he advances.

We begin by using as a dependent variable the top 1 percent share (see Alvaredo, Atkinson, Piketty, and Saez's World Top Incomes Database at http://topincomes.parisschoolofeconomics.eu/). We combine this variable with GDP data from Madison's dataset. For the first part of our analysis, we do not use explicit data on interest rates, which gives us an unbalanced panel spanning 1870–2012. For the rest of our analysis, our panel covers the post–World War II period and uses GDP data from the Penn World Tables.[4]

[4] The number of countries varies depending on the measure of the interest rate used and specification. In columns 1–3 panel A, we have 27 countries: Argentina, Australia, Canada, China, Colombia, Denmark, Finland, France, Germany, India, Indonesia, Ireland, Italy, Japan, Malaysia, Mauritius, Netherlands, New Zealand, Norway, Portugal, Singapore, South Africa, Spain, Sweden, Switzerland, United Kingdom, and United States. In column 2 panel B, we lose China and Colombia, and additionally Portugal in column 3. In column 4 panel A, we lose the non-OECD countries, China, Colombia, India, Indonesia, Malaysia, Mauritius, and Singapore relative to columns 1–3, and additionally Germany in columns 5 and 6. In panel B, we additionally lose Portugal in columns 4 and 5, and Portugal and Germany in column 6. In column 7 panel B, we have Uruguay in addition to the 27 countries in column 1. In columns 8 and 9, we lose Germany and Uruguay. In panel B, we lose Uruguay in column 7 relative to panel A, and additionally China and Colombia in column 8, and Argentina, China, Colombia, Indonesia, and Portugal in column 9.

Table 1 reports regressions using three different measures of $r - g$. First, we assume that all capital markets are open and all of the countries in the sample have the same (possibly time-varying) interest rate. Under this assumption, cross-country variation in $r - g$ will arise only because of variation in the growth rate, g. The first three columns in panel A of this table then simply exploit variation in g using annual data (that is, we set $r - g = -g$ by normalizing $r = 0$). Throughout, the standard errors are corrected for arbitrary heteroskedasticity and serial correlation at the country level; and because the number of countries is small (varying between 18 and 28), they are computed using the pairs-cluster bootstrap procedure proposed by Cameron, Gelbach, and Miller (2008), which has better finite-sample properties than the commonly used clustered standard errors. (The same results with "traditional" standard errors that assume no heteroskedasticity and residual serial correlation are reported in Appendix Table A1 and show very similar patterns.) In column 1, we look at the relationship between annual top 1 percent share and annual growth in a specification that includes a full set of year dummies and country dummies—so that the pure time-series variation at the world level is purged by year dummies and none of the results rely on cross-country comparisons. Piketty's theory predicts a positive and significant coefficient on this measure of $r - g$: that is, in countries with higher g, the incomes of the bottom 99 percent will grow more, limiting the top 1 percent share.[5] Instead, we find a negative estimate that is statistically insignificant.

In column 2, we include five annual lags of top 1 percent share on the right-hand side to model the significant amount of persistence in measures of inequality. Though specifications that include the lagged dependent variable on the right-hand side are potentially subject to the Nickell (1981) bias, given the length of the panel here this is unlikely to be an issue (since this bias disappears as the time dimension becomes large). The test at the bottom of the table shows that lagged top 1 percent share is indeed highly significant. In this case, the impact of $r - g$ is negative and significant at 10 percent—the opposite of the prediction of *Capital in the Twenty-First Century*. Column 3 includes five annual lags of GDP as well as five lags of top 1 percent share simultaneously. There is once more no evidence of a positive impact of $r - g$ on top inequality. On the contrary, the relationship is again negative, as shown by the first lag and also by the long-run cumulative effect reported at the bottom.

What matters for inequality may not be annual or five-year variations exploited in panel A, but longer-term swings in $r - g$. Panel B investigates this possibility by looking at 10-year (columns 1, 2, 4, 5, 7, 8) and 20-year data (columns 3, 6, 9).[6]

[5] With returns to capital determined in the global economy, that is, $r_{it} = r_t$ (where i refers to country and t the time period), variation in r_t is fully absorbed by the time effects in these regression models, making the $r = 0$ normalization without any loss of generality. Note, however, that what determines the dynamics of inequality in a country according to Piketty's general law is that country's growth rate, supporting the methodology here, which exploits country-specific variation in growth rates (conditional on country and time fixed effects).

[6] To avoid the mechanical serial correlation that would arise from averaging the dependent variable, we take the top 1 percent share observations every 10 or 20 years. If an observation is missing at those dates and there exists an observation within plus or minus two years, we use these neighboring observations. The results are very similar with averaging.

Table 1
Regression Coefficients of Different Proxies of $r - g$
(dependent variable is the top 1 percent share of national income)

	No cross-country variation in r			*OECD data on interest rates*			r = MPK − δ		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*
Panel A: Estimates using annual panel									
Estimate of $r - g$ at t	−0.006	−0.018*	−0.018*	−0.066**	−0.038**	−0.040*	0.029	−0.004	−0.011
	(0.012)	(0.010)	(0.011)	(0.027)	(0.017)	(0.021)	(0.033)	(0.009)	(0.008)
Estimate of $r - g$ at $t - 1$			0.001			−0.003			0.005
			(0.009)			(0.015)			(0.014)
Estimate of $r - g$ at $t - 2$			0.005			0.010			−0.012
			(0.008)			(0.019)			(0.008)
Estimate of $r - g$ at $t - 3$			−0.002			−0.012			0.014*
			(0.008)			(0.024)			(0.008)
Estimate of $r - g$ at $t - 4$			−0.005			−0.005			0.006
			(0.007)			(0.013)			(0.010)
Joint significance of lags [p-value]			4.55			7.47			12.40
			[0.47]			[0.19]			[0.03]
Long-run effect [p-value estimate > 0]		−0.16	−0.18		−0.39	−0.47		−0.04	0.03
		[0.13]	[0.15]		[0.29]	[0.34]		[0.68]	[0.89]
Persistence of top 1 percent share [p-value estimate < 1]		0.89	0.89		0.90	0.89		0.90	0.92
		[0.00]	[0.00]		[0.31]	[0.30]		[0.11]	[0.18]
Observations	1,646	1,233	1,226	627	520	470	1,162	905	860
Countries	27	27	27	19	18	18	28	26	26

(continued)

These specifications do not provide any evidence of a positive relationship between this measure of $r - g$ and top 1 percent share either.

In columns 4–6 in panel A, we work with a different measure of $r - g$ based on the realized interest rate constructed from data on nominal yields of long-term government bonds and inflation rates from the OECD. The relationship is again negative and now statistically significant at 5 percent in columns 4 and 5, and at 10 percent in column 6. In panel B, when we use 10- and 20-year panels, the relationship continues to be negative but is now statistically insignificant.

One concern with the results in columns 4–6 is that the relevant interest rate for the very rich may not be the one for long-term government bonds. Motivated by this possibility, columns 7–9 utilize the procedure proposed by Caselli and Feyrer (2007) to estimate the economy-wide marginal product of capital minus the depreciation rate using data on aggregate factors of production, and construct $r - g$ using these estimates. Now the relationship is more unstable. In some specifications it becomes positive but is never statistically significant.

Appendix Tables A2 and A3 show that these results are robust to including, additionally, GDP per capita (as another control for the business cycle and its impact on the top 1 percent share), population growth, and country-specific trends, and

Table 1—Continued

	No cross-country variation in r			*OECD data on interest rates*			$r =$ MPK $- \delta$		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*
Panel B: Estimates using 10-year (columns 1, 2, 4, 5, 7, 8) and 20-year (columns 3, 6, 9) panels									
Average $r - g$	0.055	−0.036	−0.252	−0.114	−0.121	−0.110	0.069	0.148	0.238
	(0.110)	(0.118)	(0.269)	(0.138)	(0.132)	(0.320)	(0.118)	(0.100)	(0.164)
Long-run effect		−0.05			−0.25			0.29	
[*p*-value estimate > 0]		[0.76]			[0.44]			[0.22]	
Persistence of top		0.32			0.52			0.48	
1 percent share		[0.00]			[0.02]			[0.00]	
[*p*-value estimate < 1]									
Observations	213	181	106	82	80	43	135	124	61
Countries	27	25	24	18	18	17	27	25	22

Notes: The table presents estimates of different proxies of $r - g$ on the top 1 percent share of national income. The dependent variable is available from 1871 onwards for the countries covered in the World Top Incomes Database. We use different proxies of $r - g$: Columns 1 to 3 use growth rates from Madisson, and assume no variation in real interest rates across countries. These data are available from 1870 onwards. Columns 4 to 6 use real interest rates computed by subtracting realized inflation from nominal yields on long-term government bonds, and growth rates from the Penn World Tables. These data are only available since 1955 for OECD countries. Columns 7 to 9 use $r =$ MPK $- \delta$, constructed as explained in the text using data from the Penn World Tables, and growth rates from the Penn World Tables. These data are available for 1950 onwards. Panel A uses an unbalanced yearly panel. Columns 2, 5, and 8 add five lags of the dependent variable and report the estimated persistence of the top 1 percent share of national income and the estimated long run effect of $r - g$ on the dependent variable. Columns 3, 6, and 9 add four lags of $r - g$ on the right-hand side, and also report the long-run effect of a permanent increase of 1 percent in $r - g$ and a test for the joint significance of these lags (with its corresponding χ^2 statistic and *p*-value). Panel B uses an unbalanced panel with observations every 10 years or 20 years (columns 3, 6, 9). Columns 1, 2, 4, 5, 7, and 8 present estimates from a regression of the top 1 percent share of national income at the end of each decade in the sample (that is, 1880, 1890, . . ., 2010, depending on data availability) on the average $r - g$ during the decade. Columns 2, 5, and 8 add one lag of the dependent variable on the right-hand side. Finally, columns 3, 6, and 9 present estimates from a regression of the top 1 percent share of national income at the end of each 20-year period in the sample (that is, 1890, 1910, . . ., 2010, depending on data availability) on the average $r - g$ during the period. All specifications include a full set of country and year fixed effects. Standard errors allowing for arbitrary heteroskedasticity and serial correlation of residuals at the country level are computed using the pairs-cluster bootstrap procedure proposed by Cameron, Gelbach, and Miller (2008) and are reported in parentheses.
*, **, and *** indicate 10, 5, and 1 percent levels of significance, respectively.

to the use of the top 5 percent measure of inequality as the dependent variable. Appendix Table A4 verifies that the results are similar if we limit the analysis to a common sample consisting of OECD countries since 1950, and Appendix Table A5 shows that focusing on the capital share of national income, rather than the top 1 percent share, leads to a similar set of results, providing no consistent evidence of an impact from $r - g$ to inequality.[7]

[7] This table uses two alternative measures of the capital share of national income from the Penn World Tables and from the OECD. We do not present regressions using the marginal product of capital from Caselli and Feyrer (2007) as this measure is computed using the capital share of national income, making it mechanically correlated with the dependent variable in this table.

Although this evidence is tentative and obviously we are not pretending to estimate any sort of causal relationship between $r - g$ and the top 1 percent share, it is quite striking that such basic conditional correlations provide no support for the central emphasis of *Capital in the Twenty-first Century*.[8] This is not to say that a higher r is not a force towards greater inequality in society—it probably is. It is just that there are many other forces promoting inequality and our regressions suggest that, at least in a correlational sense, these are quantitatively more important than $r - g$.

A Tale of Two Inequalities: Sweden and South Africa

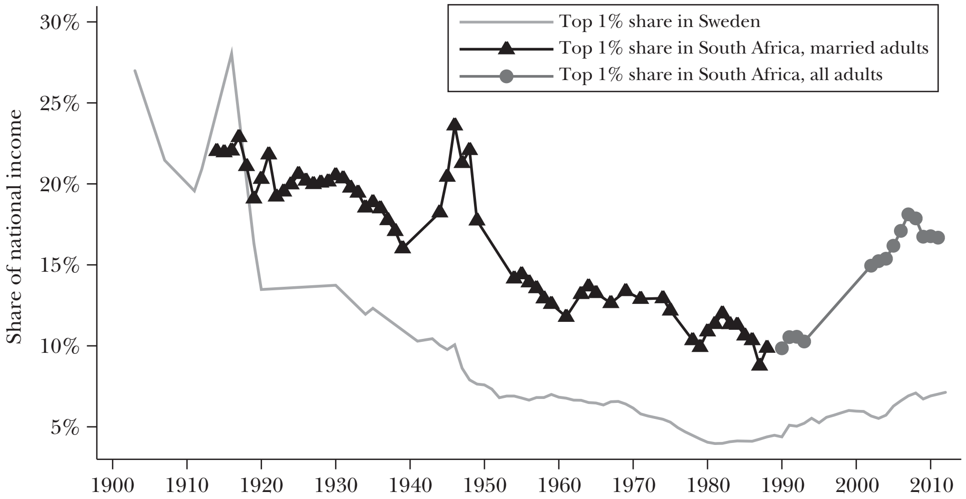
We now use the histories of inequality during the 20th century in Sweden and South Africa to illustrate how the dynamics of inequality appear linked to the institutional paths of these societies—rather than to the forces of $r > g$. In addition, these cases illustrate that the share of national income going to the top 0.1 percent or top 1 percent can give a distorted view of what is actually happening to inequality more broadly. Indeed, this focus on inequality at the top inevitably leads to a lesser and insufficient focus on what is taking place in the middle or the bottom of the income distribution.

Figure 1 shows the evolution of the share of the top 1 percent in national income in Sweden and South Africa since the early 20th century. There are of course some differences. Sweden started out with a higher top 1 percent share than South Africa, but its top 1 percent share fell faster, especially following World War I. The recent increase in the top 1 percent also starts earlier in Sweden and is less pronounced than what we see in South Africa in the 1990s and 2000s. But in broad terms, the top 1 percent share behaves similarly in the two countries, starting high, then falling almost monotonically until the 1980s, and then turning up. Such common dynamics for the top 1 percent share in two such different countries—a former colony with a history of coerced labor and land expropriation, ruled for much of the 20th century by a racist white minority, on the one hand, and the birthplace of European social democracy, on the other—would seem to bolster Piketty's case that the general laws of capitalism explain the big swings of inequality, with little reference to institutions and politics. Perhaps one could even claim, as in Piketty's example of the French Revolution, that the effects of apartheid and social democracy are trifling details against the fundamental force of $r > g$.

Except that the reality is rather different. In South Africa, for example, the institutionalization of white dominance after 1910 quickly led to the Native Land Act in 1913 which allocated 93 percent of the land to the "white economy" while

[8] One important caveat is that the ex post negative returns that may have resulted from stock market crashes and wars are not in our sample, because our estimates for r are from the post–World War II sample. Nevertheless, if $r - g$ is indeed a fundamental force towards greater inequality, we should see its impact during the last 60 years also.

Figure 1
Top 1 Percent Shares of National Income in Sweden and South Africa

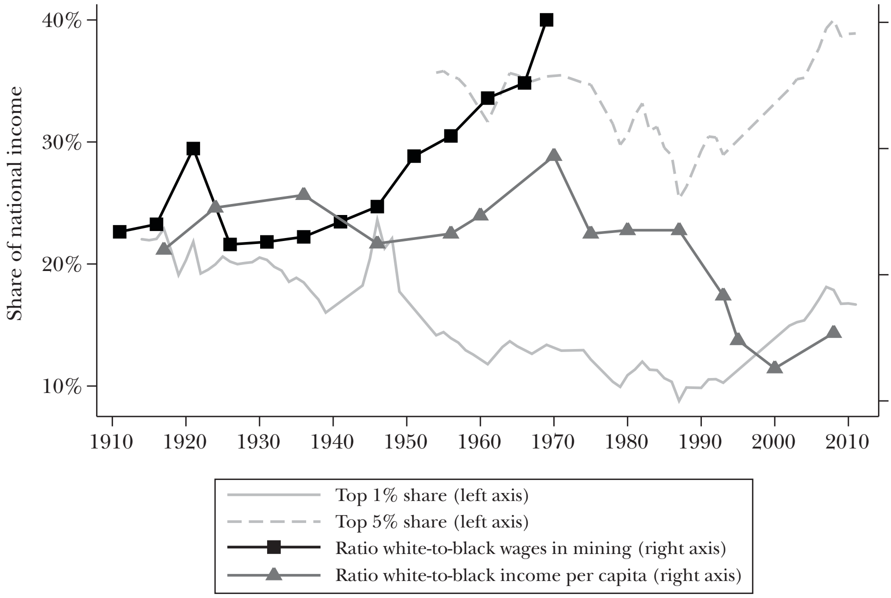


Sources: The data series for South Africa is from Alvaredo and Atkinson (2010). The data series for Sweden is from Roine and Waldenström (2009).
Note: The figure plots the top 1 percent share of national income for South Africa and Sweden.

the blacks (around 59 percent of the population) got 7 percent of the land. In the white economy, it became illegal for blacks to own property or a business, and many types of contractual relations for blacks were explicitly banned. By the 1920s, the "color bar" blocked blacks from practically all skilled and professional occupations (van der Horst 1942; Feinstein 2005, chap. 2–4). After 1948, the apartheid state became even stronger, implementing a wide array of measures to enforce social and educational segregation between whites and blacks. Finally, in 1994, the apartheid institutions collapsed as Nelson Mandela became South Africa's first black president. However, a naïve look at Figure 1 would seem to suggest that South Africa's apartheid regime, which was explicitly structured to keep black wages low and to benefit whites, was responsible for a great decrease in inequality, while the end of apartheid caused an explosion in inequality!

How can this be? The answer is that measuring inequality by the top 1 percent share can give a misleading picture of inequality dynamics in some settings. Figure 2 shows the top 1 percent share together with other measures of inequality in South Africa, which behave quite differently. Inequality between whites and blacks was massively widening during the 20th century as measured by the ratio of white-to-black wages in gold mining, a key engine of the South African economy at the time (from the wage series of Wilson 1972); this represents a continuation of 19th-century trends (discussed in de Zwart 2011). This pattern is confirmed by the white-to-black per capita income ratio from census data, which has some ups

Figure 2
Top Income Shares and Between-Group Inequality in South Africa



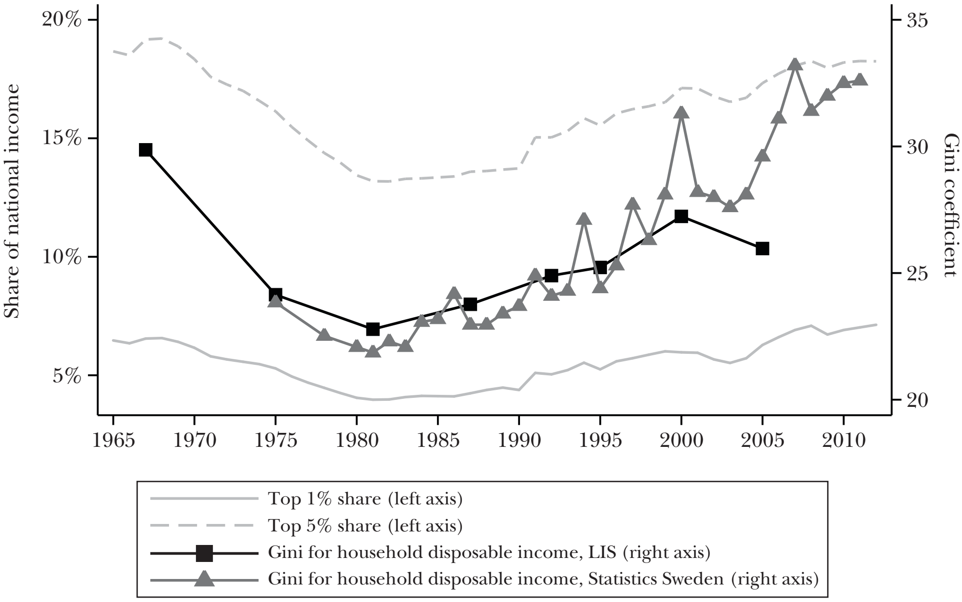
Sources and Notes: The left axis shows the top 1 and 5 percent shares of national income for South Africa on the left axis, obtained from Alvaredo and Atkinson (2010). The right axis shows the ratio between whites' and blacks' wages in mining (obtained from Wilson, 1972), and the ratio between whites' and blacks' income per capita (obtained from Leibbrandt, Woolard, Finn, and Argent 2010).

and downs but exhibits a fairly large increase from about 11-fold to 14-fold from 1911 until 1970. Thereafter, it shows a rapid decline. Even the top 5 percent share behaves somewhat differently than the top 1 percent share (though available data for this variable start only in the 1950s).

If one wanted to understand economic inequality in South Africa, changes in labor market institutions and political equilibria appear much more relevant than *r* and *g*. Indeed, the alternative measures of inequality in Figure 2 show that during the time the share of the top 1 percent was falling, South Africa became one of the most unequal countries in the world. As we will discuss, the turning points in inequality in South Africa in fact have institutional and political roots.

Figure 3 shows that in Sweden, the decline in the top 1 percent share from 1965 to 1980 is accompanied by a much more pervasive fall in inequality as measured by the Gini coefficient for household disposable income. And over the entire period, the two series for the Gini index have similar trends to the top 1 percent and the top 5 percent shares. However, in the Swedish case as well, the story of inequality seems related not to supposed general laws of capitalism and changes in *r* and *g*, but rather to institutional changes (Bengtsson 2014). The initial fall in the top 1 percent share coincided with

Figure 3
Top Income Shares and Overall Inequality in Sweden



Notes: The figure plots the top 1 and 5 percent shares of national income for Sweden on the left vertical axis, obtained from Roine and Waldenström (2009). The right axis plots the Gini coefficient for household disposable income, from the Luxembourg Income Study (Milanovic 2013), and from Statistics Sweden (SCB).

large changes in government policy: for example, a rapid increase in redistribution in the 1920s from practically nothing in the 1910s (Lindert 1994), and an increase in top marginal tax rates from around 10 percent in 1910 to 40 percent by 1930 and 60 percent by 1940 (Roine, Vlachos, and Waldenström 2009, p. 982). The expanding role of the government and of redistributive taxation plausibly had a negative impact on the top 1 percent share. The data in Figures 1 and 3 are for pre-tax inequality, but these are likely to be affected by taxes, which influence effort and investment (see the evidence in Roine, Vlachos, and Waldenström 2009), and also directly by the wage compression created by Sweden's labor market institutions. Indeed, union density rose rapidly from around 10 percent of the labor force during World War I to 35 percent by 1930 and to over 50 percent by 1940 (Donado and Wälde 2012).

Piketty emphasizes the role of the destruction of the capital stock and asset price falls in the aftermath of the two world wars as key factors explaining the decline of top inequality during much of the 20th century. But such factors can hardly account for the trends in Sweden or South Africa. Sweden was neutral in both wars, and though South Africa provided troops and resources for the Allied powers in both, neither economy experienced any direct destruction of their capital stock.

Towards an Institutional Framework

A satisfactory framework for the analysis of inequality should take into account both the effect of different types of institutions on the distribution of resources and the endogenous evolution of these institutions. We now flesh out such a framework and then apply it to the evolution of inequality—and institutions—in Sweden and South Africa. The framework we present is based on the one we proposed in Acemoglu, Johnson, and Robinson (2005). Adapting Figure 1 from that paper, our framework can be represented schematically as follows:

$$\begin{array}{c}\text{political}\\ \text{institutions}_t \Longrightarrow\\[3em] \text{inequality}_t \Longrightarrow\end{array}
\left.\begin{array}{c}\text{de jure}\\ \text{political}\\ \text{power}_t\\ \&\\ \text{de facto}\\ \text{political}\\ \text{power}_t\end{array}\right\}
\begin{array}{l}\Longrightarrow \begin{array}{c}\text{economic}\\ \text{institutions}_t\end{array} \Longrightarrow \left.\begin{array}{c}\text{technology}_t,\\ \text{skills}_t, \ \&\\ \text{prices}_t\end{array}\right\} \Longrightarrow \left\{\begin{array}{c}\text{economic}\\ \text{performance}_t\\ \&\\ \text{inequality}_{t+1}\end{array}\right.\\[3em] \Longrightarrow \begin{array}{c}\text{political}\\ \text{institutions}_{t+1}\end{array}\end{array}$$

In this approach, the prevailing political institutions at a certain time determine the distribution of de jure political power (Acemoglu and Robinson 2000, 2008; Acemoglu 2008; Acemoglu, Egorov, and Sonin 2012, forthcoming): for example, which groups are disenfranchised, how political power is contested, how constrained the economic and political elites are, and so on. Political institutions also affect, together with inequality in society, the distribution of de facto political power. For instance, de facto power—which designates political power and constraints generated by access to the means of violence, collective action, informal institutions, and social norms—depends on the extent to which different social and economic groups are organized and how they resolve their collective action problems and how resources influence their ability to do so. De facto and de jure power together determine economic institutions and also the stability and change of political institutions.

In turn, economic institutions affect the supply of skills—a crucial determinant of inequality throughout history and even more so today. Economic institutions also, through regulation of both prices and market structure, by taxation, or by affecting the bargaining power of different factors of production and individuals, influence goods and factor prices. Finally, economic institutions affect technology, including whether and how efficiently existing technologies are utilized, as well as the evolution of technology through endogenous innovations and learning by doing. For example, Zeira (1998) and Acemoglu (2010) show how low wages, resulting from either supply or institutional factors, can sometimes reduce technology adoption or even technological progress, and Hornbeck and Naidu (2014) provide evidence consistent with this pattern. Through their joint impact on technology, the supply of skills, and relative prices, economic institutions affect not only r and g, but more importantly, inequality. In this approach, inequality should not be thought of as always summarized by a single statistic, such as the Gini index or the top 1 percent share. Rather, the economic and political factors stressed here determine the distribution of resources more generally.

We do not mean to suggest that this framework determines the evolution of institutions, technology, and inequality deterministically. The arrows designate influences, which are mediated by various stochastic events and political economy interactions, and similar economic developments will result in very different institutional responses depending on the prevailing political equilibrium, as evidenced by the contrasting histories of Mexico and the United States in the 20th century (noted earlier). Nor do we imply that the framework captures all economic implications of import—or all of those that are relevant for inequality. Most centrally, technology will evolve over time not only because of institutional factors, but also due to scientific developments and because it responds to other economic changes, including factor prices, the abundance and scarcity of different types of skills and market structure (for example, Acemoglu 2002, 2003, 2010). It is possible as well that technological developments could in turn affect institutional dynamics (for example, Acemoglu, Aghion, and Violante 2001; Hassler, Rodriguez Mora, Storlesletten, and Zilibotti 2003). Nevertheless, this simple framework is useful for highlighting the potentially important role of institutional equilibria, and their changes, in shaping inequality.

Let us now apply it to South Africa. Before 1910, non-whites could vote in the Cape and Natal as long as they fulfilled certain wealth, income, or property restrictions (though this was more heavily restricted in Natal). After 1910, a specifically white franchise was established in the Transvaal and Orange Free State, and then gradually extended to the rest of the country with blacks finally being definitively disenfranchised in the Cape in 1936. The de jure institutions of the apartheid state cemented the political power of the white minority, and segregationist laws and other aspects of the regime created economic institutions, such as the skewed distribution of land and the "color bar," aimed at furthering the interests of the white minority. So then why did this and the flourishing of social apartheid after 1948 lead to a fall in the top 1 percent share?

The primary reason is that political dynamics in South Africa at this time cannot be fully captured as a conflict between monolithic groups of whites and blacks. Rather, apartheid should be viewed as a coalition between white workers, farmers, and mine-owners—at the expense of blacks but also white industrialists who had to pay very high wages for white workers (Lundahl 1982; Lipton 1985). Thus, one reason for a reduction in the top 1 percent share was that profits were squeezed by wages for white labor. Moreover, by depriving industrialists of a larger pool of skilled workers, and tilting the price of white labor higher (because the supply of labor was artificially restricted), these rules further stunted South African economic development.

In addition, there were forces within apartheid for redistribution from the very rich towards poorer whites. Indeed, South Africa's political discussions in the 1920s that led to the further spread of the "color bar" and subsequently to the victory of the National Party in 1948 were related to what was called the "poor white problem," highlighting the importance of the specific coalition underpinning apartheid. Alvaredo and Atkinson (2010) discuss other factors such as the gold price.

The compression of the huge wage gaps between South Africa's whites and blacks starting in the 1970s (see Figure 2) should be viewed within the context of the political weakening of the apartheid regime and its increasing economic problems (Wilson 1980; Mariotti 2012). The domestic turning point was the ability of black workers to organize protests and riots, and exercise their de facto power, particularly after the Soweto uprising of 1976, which led to the recognition of black trade unions. This process was aided by mounting international pressure, which induced British and US firms based in South Africa to push back against workplace discrimination. Ultimately, this de facto power forced the collapse of the apartheid regime, leading to a new set of political institutions and the enfranchisement of black South Africans. The new set of economic institutions, and their consequences for inequality, flowed from these political changes. Consistent with our framework, the institutions of apartheid may have also fed back into the evolution of technology, for example in impeding the mechanization of gold mining (Spandau 1980). As the power of apartheid started to erode in the 1970s, white businessmen responded rapidly by substituting capital for labor and moving technology in a labor-saving direction (Seekings and Nattrass 2005, p. 403).

As can be seen from Figure 1, the top 1 percent share in South Africa shows a steep rise after 1994, coinciding with the final overthrow of the formidable extractive institutions of apartheid. No clear consensus has yet emerged on the causes of the post-apartheid increase in inequality, but one reason relates to the fact that after the end of apartheid, the artificially compressed income distribution of blacks started widening as some portion of the population started to benefit from new business opportunities, education, and aggressive affirmative action programs (Leibbrandt, Woolard, Finn, and Argent 2010). Whatever the details of these explanations, it is hard to see the post-1994 rise in the top 1 percent share as representing the demise of a previously egalitarian South Africa.

The role of de facto and de jure political power in shaping political and economic institutions is no less central in Sweden, where the important turning point was created by the process of democratization. Adult male suffrage came in 1909, but true parliamentary democracy developed only after the Reform Act of 1918, with significant curbs on the power of the monarchy and more competitive elections. Both the 1909 reform and the emergence of parliamentary democracy in 1918 were responses to unrest, strikes, and the de facto power of disenfranchised workers, especially in the atmosphere of uncertainty and social unrest following World War I (Tilton 1974). Collier (1999, p. 83) explains: "[I]t was only after the economic crisis of 1918 and ensuing worker protests for democracy led by Social Democrats that the Reform Act was passed. Indeed, in November 1918, labor protests reached such a point as to be perceived as a revolutionary threat by Sweden's Conservative Party and upper classes."

Swedish democracy then laid the foundations for modern labor market institutions and the welfare state, and created powerful downward pressure on inequality, including the top 1 percent share. However, democratic conflict in Sweden was not a simple contest between monolithic groups of workers and businesses either. As Moene and Wallerstein (1995, 2006) characterize it, social democracy was a coalition of the ends of the income distribution—businessmen and unskilled workers—against the middle class and skilled workers (for theories about the emergence of such political coalitions, see also Saint-Paul 2000; Gourevitch 1986; Luebbert 1991). In consequence, Swedish economic institutions strongly compressed skilled wages relative to unskilled wages, underpinning the rapid decline in broad-based measures of inequality. Some businesses benefitted from these arrangements, particularly those in sectors exposed to international competition, which used centralized wage bargaining as a tool to stop wage push from nontraded sectors, such as construction (Swenson 1991, 2002). Swedish labor market institutions also likely affected the path of technology. For instance, Moene and Wallerstein (1997) emphasize that wage compression acted as a tax on inefficient plants and stimulated new entry and rapid technological upgrading. In the face of high unskilled wages and the institutions of the welfare state, it is not a surprise that the top 1 percent share declined in Sweden as well, even if businessmen also did well with some aspects of Swedish labor market institutions.

What explains the fact that the top 1 percent share appears to increase not just in South Africa and Sweden, but in almost all OECD economies over the last 20 years or so? Factors left out of our framework—globalization, skill-biased technological changes, and the increase in the size of large corporations—are likely to be important. But these forces are themselves not autonomous but have likely responded to other changes in the world economy. For example, Acemoglu (2002) argues that skill-biased technological change cannot be understood without the increase in the supply of skilled workers in the United States and the world economy, making these types of technologies more profitable; and globalization and the increasing size of global corporations are themselves consequences of regulatory and technological changes of the last several decades. This simply underscores that the framework presented here cannot capture the dynamics of all dimensions of inequality—or the rich dynamics of political and economic institutions for that matter. Nevertheless, the basic forces that it stresses appear to be important not just in the context of Sweden and South Africa, but much more generally (as we argue in Acemoglu and Robinson 2006, 2012).

This framework also helps to clarify the reasons why we might care about inequality at the very top of the income and wealth distributions. Most relevant is that the factors undergirding a high share of income for the top 1 percent might also represent a lack of equality of opportunity or a lack of a level playing field. Extending the framework presented above, we argued in Acemoglu and Robinson (2012) that lack of a level playing field, including limited social mobility, is likely to hold back countries in their investments, innovation, and the efficiency of resource allocation. However, the top 1 percent share may not be the most relevant dimension of the distribution of income for evaluating equality of opportunity and barriers to the efficient allocation of talent and resources in society. For example, if a small number at the top became wealthier—say, if Bill Gates and Warren Buffett became twice as wealthy—at the expense of other rich individuals, would that make US society notably less meritocratic? This seems unlikely. Indeed, Chetty, Hendren, Kline, and Saez (2014) and Chetty, Hendren, Kline, Saez, and Turner (2014) show that social mobility at the commuting zone level in the United States is unrelated to income inequality, especially inequality at the top. Their evidence that US social mobility has stayed the same even as the top 1 percent share has increased rapidly over the last several decades further corroborates this intuition. Other types of inequalities, such as the gap between whites and blacks as in South Africa or between the bottom and the middle class in the United States, may be more relevant for thinking about whether there have been changes in social mobility and the angle of the playing field.

But one dimension of political economy where the top 1 percent share may be central is the health of political institutions. It may be difficult to maintain political institutions that create a dispersed distribution of political power and political access for a wide cross-section of people in a society in which a small number of families and individuals have become disproportionately rich. A cautionary tale about the dangers created by this type of inequality is discussed in Puga and Trefler (2014) and Acemoglu and Robinson (2012): the story of late medieval Venice. Here, the economic power of the most prosperous and well-established families ultimately made it possible for them to block the access of others to political power, and once they thus monopolized political power, they could change economic institutions for their benefit by blocking the entry of other families into lucrative businesses and banning contracts that had previously made it possible for individuals with limited capital to enter into partnerships for long-distance trade. This change in political institutions, feeding into a deterioration of economic institutions, heralded the economic decline of Venice.

Yet if the primary threat from the top 1 percent share is political, then the main response should be related to monitoring and containing the political implications of the increase in top-level inequality—not necessarily catch-all policies such as the wealth taxes advocated by Piketty. Such policies should be explicitly related to the institutional fault lines of the specific society and should be conceived in the context of strengthening institutional checks against any potential power grab.

Conclusion

Thomas Piketty's (2014) ambitious work proffers a bold, sweeping theory of inequality applicable to all capitalist economies. Though we believe that the focus on inequality and the ensuing debates on policy are healthy and constructive, we have argued that Piketty goes wrong for exactly the same reasons that Karl Marx, and before him David Ricardo, went astray. These quests for general laws ignore both institutions and politics, and the flexible and multifaceted nature of technology, which make the responses to the same stimuli conditional on historical, political, institutional, and contingent aspects of the society and the epoch, vitiating the foundations of theories seeking fundamental, general laws. We have argued, in contradiction to this perspective, that any plausible theory of the nature and evolution of inequality has to include political and economic institutions at the center stage, recognize the endogenous evolution of technology in response to both institutional and other economic and demographic factors, and also attempt to model how the response of an economy to shocks and opportunities will depend on its existing political and institutional equilibrium.

■ *We thank David Autor, Amy Finkelstein, Johan Fourie, Bengt Holmstrom, Chang Tai Hsieh, Chad Jones, Matthew Kustenbauder, Naomi Lamoureux, Ulrike Malmendier, Kalle Moene, Joel Mokyr, Suresh Naidu, Jim Poterba, Matthew Rognlie, Ragnar Torvik, Laurence Wilse-Samson, Francis Wilson, and Timothy Taylor for their comments and Pascual Restrepo for extensive discussions, comments, and superb research assistance.*

References

Acemoglu, Daron. 2002. "Directed Technical Change." *Review of Economic Studies* 69(4): 781–809.

Acemoglu, Daron. 2003. "Labor- and Capital-Augmenting Technical Change." *Journal of the European Economic Association* 1(1): 1–37.

Acemoglu, Daron. 2008. "Oligarchic versus Democratic Societies." *Journal of the European Economic Association* 6(1): 1–44.

Acemoglu, Daron. 2010. "When Does Labor Scarcity Encourage Innovation?" *Journal of Political Economy* 118(6): 1037–78.

Acemoglu, Daron, Philippe Aghion, and Giovanni L. Violante. 2001. "Deunionization, Technical Change and Inequality." *Carnegie-Rochester Conference Series on Public Policy* 55(1): 229–64.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics Volume 4B*, edited by Orley Ashenfelter and David Card, 1043–171. Elsevier, North-Holland.

Acemoglu, Daron, Davide Cantoni, Simon Johnson, and James A. Robinson. 2011. "The Consequences of Radical Reform: The French Revolution." *American Economic Review* 101(7): 3286–3307.

Acemoglu, Daron, Georgy Egorov, and Konstantin Sonin. 2012. "Dynamics and Stability of Constitutions, Coalitions, and Clubs." *American Economic Review* 102(4): 1446–76.

Acemoglu, Daron, Georgy Egorov, and Konstantin Sonin. Forthcoming. "Political Economy in a Changing World." *Journal of Political Economy.*

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2005. "Institutions as a Fundamental Cause of Long-Run Growth." In *Handbook of Economic Growth Volume 1A*, edited by Philippe Aghion and Steven N. Durlauf, 385–472. Elsevier, North-Holland.

Acemoglu, Daron, and James A. Robinson. 2000. "Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective." *Quarterly Journal of Economics* 115(4): 1167–99.

Acemoglu, Daron, and James A. Robinson. 2006. *Economic Origins of Dictatorship and Democracy.* New York: Cambridge University Press.

Acemoglu, Daron, and James A. Robinson. 2008. "Persistence of Power, Elites, and Institutions." *American Economic Review* 98(1): 267–93.

Acemoglu, Daron, and James A. Robinson.

2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York: Crown.

Aiyagari, S. Rao. 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving." *Quarterly Journal of Economics* 109(3): 659–84.

Allen, Robert C. 2001. "The Great Divergence in European Wages and Prices from the Middle Ages to the First World War." *Explorations in Economic History* 38(4): 411–47.

Allen, Robert C. 2007. "Pessimism Preserved: Real Wages in the British Industrial Revolution." Oxford University Department of Economics Working Paper 314. http://www.nuffield.ox.ac.uk/users/Allen/unpublished/pessimism-6.pdf.

Allen, Robert C. 2009a. "Engels' Pause: Technological Change, Capital Accumulation, and Inequality in the British Industrial Revolution." *Explorations in Economic History* 46(4): 418–35.

Allen, Robert C. 2009b. *The British Industrial Revolution in Global Perspective.* New York: Cambridge University Press.

Alvaredo, Facundo, and Anthony B. Atkinson. 2010. Colonial Rule, Apartheid and Natural Resources: Top Incomes in South Africa, 1903–2007. Centre for Economic Policy Research Discussion Paper 8155. http://www.parisschoolofeconomics.eu/IMG/pdf/DP8155_SouthAfrica.pdf.

Alvaredo, Facundo, Tony Atkinson, Thomas Piketty, and Emmanuel Saez, compilers. World Top Incomes Database, http://topincomes.parisschoolofeconomics.eu/.

Antràs, Pol. 2004. "Is the U.S. Aggregate Production Function Cobb–Douglas? New Estimates of the Elasticity of Substitution." *Contributions to Macroeconomics* 4(1): Article 4.

Bengtsson, Erik. 2014. "Labour's Share in Twentieth Century Sweden: A Reinterpretation." *Scandinavian Economic History Review* 62 (3):290–314.

Bewley, Truman F. 1986. "Stationary Monetary Equilibrium with a Continuum of Independently Fluctuating Consumers." In *Contributions to Mathematical Economics in Honor of Gérard Debreu*, edited by Werner Hildenbrand and Andreu Mas-Colell, 79–102. Elsevier North-Holland.

Blaug, Mark. 1996. *Economic Theory in Retrospect,* 5th Edition. Cambridge University Press.

Bloch, Marc. 1967. *Land and Work in Medieval Europe: Selected Papers.* Translated by J. E. Anderson. New York: Harper Torchbooks.

Bonnet, Odran, Pierre-Henri Bono, Guillaume Chapelle, and Étienne Wasmer. 2014. "Does Housing Capital Contribute to Inequality? A Comment on Thomas Piketty's Capital in the 21st Century." SciencesPo Discussion Paper 2014-07.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90(3): 414–27.

Caselli, Francesco, and James Feyrer. 2007. "The Marginal Product of Capital." *Quarterly Journal of Economics* 122(2): 535–68.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *Quarterly Journal of Economics* 129(4): 1553–623.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. 2014. "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility." *American Economic Review* 104(5): 141–47.

Chirinko, Robert S. 1993. "Business Fixed Investment Spending: Modeling Strategies, Empirical Results, and Policy Implications." *Journal of Economic Literature* 31(4): 1875–1911.

Chirinko, Robert S. 2008. "σ: The Long and Short of It." *Journal of Macroeconomics* 30(2): 671–86.

Chirinko, Robert S., Steven M. Fazzari, and Andrew P. Meyer. 1999. "How Responsive Is Business Capital Formation to Its User Cost?: An Exploration with Micro Data." *Journal of Public Economics* 74(1): 53–80.

Chirinko, Robert S., and Debdulal Mallick. 2014. "The Substitution Elasticity, Factor Shares, Long-Run Growth, and the Low-Frequency Panel Model." https://ideas.repec.org/p/ces/ceswps/_4895.html.

Clark, Gregory. 2002. "Land Rental Values and Agrarian History: England and Wales, 1500–1912." *European Review of Economic History* 6(3): 281–308.

Clark, Gregory. 2005. "The Condition of the Working Class in England, 1209–2004." *Journal of Political Economy* 113(6): 1307–40.

Clark, Gregory. 2010. "The Macroeconomic Aggregates for England, 1209–2008." In *Research in Economic History Volume 27,* edited by Alexander J. Field, 51–140. Emerald Publishing Group.

Collier, Ruth B. 1999. *Paths toward Democracy: Working Class and Elites in Western Europe and South America.* New York: Cambridge University Press.

Collins, Norman R. and Lee E. Preston. 1961. "The Size Structure of the Largest Industrial Firms, 1909–1958." *American Economic Review* 51(5): 986–1011.

Crafts, N. F. R. 1985. *British Economic Growth during the Industrial Revolution.* Oxford: Clarendon Press.

de Zwart, Pim. 2011. "South African Living Standards in Global Perspective, 1835–1910." *Economic History of Developing Regions* 26(1): 49–74.

Diamond, Peter A. 1965. "National Debt in a Neoclassical Growth Model." *American Economic Review* 55(5): 1126–50.

Donado, Alejandro, and Klaus Wälde. 2012. "How Trade Unions Increase Welfare." *Economic Journal* 122(563): 990–1009.

Edwards, Richard C. 1975. "Stages in Corporate Stability and the Risks of Corporate Failure." *Journal of Economic History* 35(2): 428–57.

Feinstein, Charles H. 1998. "Pessimism Perpetuated: Real Wages and the Standard of Living in Britain During and After the Industrial Revolution." *Journal of Economic History* 58(3): 625–58.

Feinstein, Charles H. 2005. *An Economic History of South Africa: Conquest, Discrimination and Development.* New York: Cambridge University Press.

Foley, Duncan K. 2008. *Adam's Fallacy: A Guide to Economic Theology.* Cambridge: Belknap Press.

Geerolf, François. 2013. "Reassessing Dynamic Efficiency." https://dl.dropboxusercontent.com/u/7363883/Efficiency_Emp.pdf.

Gourevitch, Peter. 1986. *Politics in Hard Times: Comparative Responses to International Economic Crises.* New York: Cornell University Press.

Hamermesh, David S. 1993. *Labor Demand.* Princeton: Princeton University Press.

Hassler, Jon, José V. Rodríguez Mora, Kjetil Storesletten, and Fabrizio Zilibotti. 2003. "The Survival of the Welfare State." *American Economic Review* 93(1): 87–112.

Hornbeck, Richard, and Suresh Naidu. 2014. "When the Levee Breaks: Black Migration and Economic Development in the American South." *American Economic Review* 104(3): 963–90.

Karabarbounis, Loukas, and Brent Nieman. 2014. "The Global Decline of the Labor Share." *Quarterly Journal of Economics* 129(1): 61–103.

Katz, Lawrence F., and David H. Autor. 1999. "Changes in the Wage Structure and Earnings Inequality." In *Handbook of Labor Economics Volume 3A*, edited by Orley C. Ashenfelter and David Card, 1463–555. Elsevier, North-Holland.

Klump, Rainer, Peter McAdam, and Alpo Willman. 2007. "Factor Substitution and Factor–Augmenting Technical Progress in the United States: A Normalized Supply-Side System Approach." *Review of Economics and Statistics* 89(1): 183–92.

Krusell, Per, Lee E. Ohanian, José-Víctor Ríos-Rull, and Giovanni L. Violante. 2000. "Capital-Skill Complementary and Inequality: A Macroeconomic Analysis." *Econometrica* 68(5): 1029–53.

Krusell, Per, and Anthony A. Smith, Jr. 2014. "Is Piketty's 'Second Law of Capitalism' Fundamental?" http://aida.wss.yale.edu/smith/piketty1.pdf.

Lamoreaux, Naomi R. 1985. *The Great Merger Movement in American Business, 1895–1904.* New York: Cambridge University Press.

Leibbrandt, Murray, Ingrid Woolard, Arden Finn, and Jonathan Argent. 2010. "Trends in South African Income Distribution and Poverty since the Fall of Apartheid." OECD Social, Employment and Migration Working Papers, No. 101. OECD Publishing.

Lindert, Peter H. 1994. "The Rise of Social Spending, 1880–1930." *Explorations in Economic History* 31(1) 1–37.

Lipton, Merle. 1985. *Capitalism and Apartheid: South Africa, 1910–84.* London: Gower/Maurice Temple Smith.

Luebbert, Gregory M. 1991. *Liberalism, Fascism, or Social Democracy: Social Classes and the Political Origins of Regimes in Interwar Europe.* New York: Oxford University Press.

Lundahl, Mats. 1982. "The Rationale of Apartheid." *American Economic Review* 72(5): 1169–79.

Mairesse, Jacques, Bronwyn H. Hall, and Benoit Mulkay. 1999. "Firm-Level Investment in France and the United States: An Exploration of What We Have Learned in Twenty Years." *Annales d'Economie et de Statistique* 55–56(December): 27–67.

Mariotti, Martine. 2012. "Labour Markets during Apartheid in South Africa." *Economic History Review* 65(3): 1100–22.

Marx, Karl. 1847. *The Poverty of Philosophy.*

Marx, Karl. 1859. *A Contribution to the Critique of Political Economy.*

Marx, Karl. 1867. *Capital: Critique of Political Economy,* vols I, II, and III.

Mathews, R. C. O., C. H. Feinstein, and J. C. Odling-Smee. 1982. *British Economic Growth, 1856–1973: The Post-War Period in Historical Perspective.* Stanford University Press.

McLellan, David, ed. 2000. *Karl Marx: Selected Writings.* 2nd Edition. New York: Oxford University Press.

Milanovic, Branko. 2013. All the Ginies Dataset. The World Bank.

Moene, Karl Ove, and Michael Wallerstein. 1995. "How Social Democracy Worked: Labor-Market Institutions." *Politics and Society* 23(2): 185–211.

Moene, Karl Ove, and Michael Wallerstein. 1997. "Pay Inequality." *Journal of Labor Economics* 15(3): 403–30.

Moene, Karl Ove, and Michael Wallerstein. 2006. "Social Democracy as a Development Strategy." Chap. 6 in *Globalization and Egalitarian Redistribution*, edited by Pranab Bardhan, Samuel Bowles, and Michael Wallerstein. Princeton University Press/Russell Sage Foundation.

Mokyr, Joel. 2012. *The Enlightened Economy: An Economic History of Britain 1700–1850.* New Haven: Yale University Press.

Morrisson, Christian, and Wayne Snyder. 2000. "The Income Inequality of France in Historical

Perspective." *European Review of Economic History* 4(1): 59–83.

Naidu, Suresh, and Noam Yuchtman. 2013. "Coercive Contract Enforcement: Law and the Labor Market in Nineteenth Century Industrial Britain." *American Economic Review* 103(1): 107–44.

Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49(6): 1417–26.

Oberfield, Ezra, and Devesh Raval. 2014. "Micro Data and Macro Technology." Unpublished paper, Princeton University. http://economics.mit.edu/files/9861.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century.* Harvard University Press.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118(1): 1–39.

Puga, Diego, and Daniel Trefler. 2014. "International Trade and Institutional Change: Medieval Venice's Response to Globalization." *Quarterly Journal of Economics* 129(2): 753–821.

Ray, Debraj. 2014. "Nit-Piketty: A Comment on Thomas Piketty's *Capital in the Twenty First Century.*" http://www.econ.nyu.edu/user/debraj/.

Ricardo, David. 1817. *On the Principles of Political Economy and Taxation,* 1st edition.

Rognlie, Matthew. 2014. "A Note on Piketty and Diminishing Returns to Capital." http://www.mit.edu/~mrognlie/piketty_diminishing_returns.pdf.

Roine, Jesper, Jonas Vlachos, and Daniel Waldenström. 2009. "The Long-Run Determinants of Inequality: What Can We Learn from Top Income Data?" *Journal of Public Economics* 93(7–8): 974–88.

Roine, Jesper, and Daniel Waldenström. 2009. "Wealth Concentration over the Path of Development: Sweden, 1873-2006." *Scandinavian Journal of Economics* 111(1): 151-187.

Saint-Paul, Gilles. 2000. *The Political Economy of Labour Market Institutions.* New York: Oxford University Press.

Samuelson, Paul A. 1958. "An Exact Consumption-Loan Model of Interest With or Without the Social Contrivance of Money." *Journal of Political Economy* 66(6): 467–82.

Sanders, Elizabeth. 1999. *Roots of Reform: Farmers, Workers, and the American State, 1877–1917.* Chicago: University of Chicago Press.

Schonhardt-Bailey, Cheryl. 2006. *From the Corn Laws to Free Trade: Interests, Ideas, and Institutions in Historical Perspective.* Cambridge: MIT Press.

Seekings, Jeremy, and Nicoli Nattrass. 2005. *Class, Race, and Inequality in South Africa.* New Haven: Yale University Press.

Singer, Peter. 2000. *Marx: A Very Short Introduction.* New York: Oxford University Press.

Spandau, A. 1980. "Mechanization and Labour Policies on South African Mines." *South African Journal of Economics* 48(2): 110–20.

Swenson, Peter. 1991. "Bringing Capital Back In, or Social Democracy Reconsidered: Employer Power, Cross-Class Alliances, and Centralization of Industrial Relations in Denmark and Sweden." *World Politics* 43(4): 513–44.

Swenson, Peter A. 2002. *Capitalists against Markets: The Making of Labor Markets and Welfare States in the United States and Sweden.* New York: Oxford University Press.

Tilton, Timothy A. 1974. "The Social Origins of Liberal Democracy: The Swedish Case." *American Political Science Review* 68(2): 561–71.

Turner, M. E., J. V. Beckett, and B. Afton. 1999. *Agricultural Rent in England, 1690–1914.* New York: Cambridge University Press.

van der Horst, Shelia T. 1942. *Native Labour in South Africa.* London: Frank Cass and Co., Ltd.

White, Lawrence J. 1981. "What Has Been Happening to Aggregate Concentration in the United States?" *Journal of Industrial Economics* 29(3): 223–30.

White, Lawrence J. 2002. "Trends in Aggregate Concentration in the United States." *Journal of Economic Perspectives* 16(4): 137–60.

Wilson, Francis. 1972. *Labour in the South African Gold Mines, 1911–1969.* New York: Cambridge University Press.

Wilson, Francis. 1980. "Current Labor Issues in South Africa." Chap. 4 in *The Apartheid Regime: Political Power and Racial Domination*, edited by Robert M. Price and Carl G. Rosberg. University of California Press.

Zeira, Joseph. 1998. "Workers, Machines, and Economic Growth." *Quarterly Journal of Economics* 113(4): 1091–117.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 29–46

Pareto and Piketty: The Macroeconomics of Top Income and Wealth Inequality[†]

Charles I. Jones

Since the early 2000s, research by Thomas Piketty and Emmanuel Saez (and their coauthors, including Anthony Atkinson and Gabriel Zucman) has revolutionized our understanding of income and wealth inequality. The crucial point of departure for this revolution is the extensive data they have used, based largely on administrative tax records. Piketty's (2014) *Capital in the Twenty-First Century* is the latest contribution in this line of work, especially with the new data it provides on capital and wealth. Piketty also proposes a framework for describing the underlying forces that affect inequality and wealth, and unlikely as it seems, a bit of algebra that plays an important role in Piketty's book has even been seen on T-shirts: $r > g$.

In this paper, I highlight some key empirical facts from this research and describe how they relate to macroeconomics and to economic theory more generally. One of the key links between data and theory is the Pareto distribution. The paper explains simple mechanisms that give rise to Pareto distributions for income and wealth and considers the economic forces that influence top inequality over time and across countries.

To organize what follows, recall that GDP can be written as the sum of "labor income" and "capital income." This split highlights several kinds of inequality that we can explore. In particular, there is "*within*-inequality" for each of these components: How much inequality is there within labor income? How much inequality within capital income—or more appropriately here, among the wealth itself for which capital income is just the annual flow? There is also "*between*-inequality" related to the

■ *Charles I. Jones is the STANCO 25 Professor of Economics, Graduate School of Business, Stanford University, Stanford, California, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is chad.jones@stanford.edu.*

[†] To access the Appendix and Data Appendix, visit
http://dx.doi.org/10.1257/jep.29.1.29

doi=10.1257/jep.29.1.29

split of GDP between capital and labor. This between-inequality takes on particular relevance given the within-inequality fact that most wealth is held by a small fraction of the population; anything that increases between-inequality therefore is very likely to increase overall inequality.[1] In the three main sections of this paper, I consider each of these concepts in turn. I first highlight some of the key facts related to each type of inequality. Then I use economic theory to shed light on these facts.

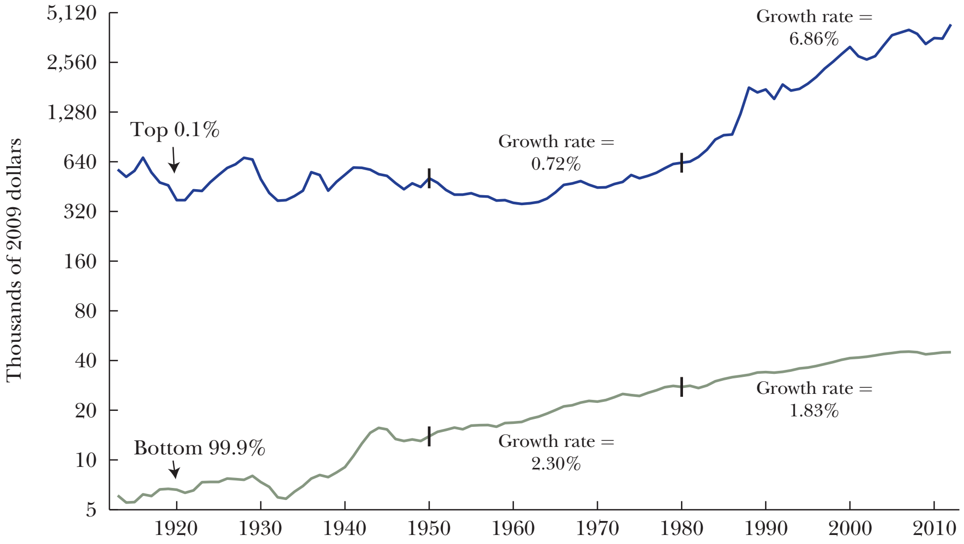
The central takeaway of the analysis is summarized by the first part of the title of the paper, "Pareto and Piketty." In particular, there is a tight link between the share of income going to the top 1 percent or top 0.1 percent and the key parameter of a Pareto distribution. Understanding why top inequality takes the form of a Pareto distribution and what economic forces can cause the key parameter to change is therefore central to understanding the facts. As just one example, the central role that Piketty assigns to $r - g$ has given rise to some confusion, in part because of its familiar presence in the neoclassical growth model, where it is not obviously related to inequality. The relationship between $r - g$ and inequality is much more easily appreciated in models that explicitly generate Pareto wealth inequality.

Capital in the Twenty-First Century, together with the broader research agenda of Piketty and his coauthors, opens many doors by assembling new data on top income and wealth inequality. The theory that Piketty develops to interpret these data and make predictions about the future is best viewed as a first attempt to make sense of the evidence. Much like Marx, Piketty plays the role of provocateur, forcing us to think about new ideas and new possibilities. As I explain below, the extent to which $r - g$ is the fundamental force driving top wealth inequality, both in the past and in the future, is unclear. But by encouraging us to entertain these questions and by providing a rich trove of data in which to study them, Piketty and his coauthors have made a tremendous contribution.

Before we begin, it is also worth stepping back to appreciate the macroeconomic consequences of the inequality that Piketty and his coauthors write about. For example, consider Figure 1. This figure is constructed by merging two famous data series: one is the Alvaredo–Atkinson–Piketty–Saez top income inequality data (about which we'll have more to say shortly) and the other is the long-run data on GDP per person for the United States that comes from Angus Maddison (pre-1929) and from the US Bureau of Economic Analysis. To set the stage, be aware that GDP per person since 1870 looks remarkably similar to a straight line when plotted on a log scale, exhibiting a relatively constant average growth rate of around 2 percent per year. Figure 1 applies the Piketty–Saez inequality shares to average GDP per person to produce an estimate of GDP per person for the top 0.1 percent and another for the bottom 99.9 percent. It is important to note that this estimate is surely imperfect. GDP likely does not follow precisely the same distribution as "adjusted gross income" in the income tax data: health insurance benefits are more equally distributed, for example. However, even with these caveats, the estimate still seems useful.

[1] One could also productively explore the correlation of the two *within* components: Are people at the top of the labor income distribution also at the top of the capital income and wealth distributions?

Figure 1
GDP per Person, Top 0.1 Percent and Bottom 99.9 Percent



Sources: Aggregate GDP per person data are taken from the Bureau of Economic Analysis (since 1929) and Angus Maddison (pre-1929). The top income share used to divide the GDP is from the October 2013 version of the World Top Incomes Database (Alvaredo, Atkinson, Piketty, and Saez n.d.).
Notes: This figure displays an estimate of average GDP per person for the top 0.1 percent and the bottom 99.9 percent. Average annual growth rates for the periods 1950–1980 and 1980–2007 are also reported.

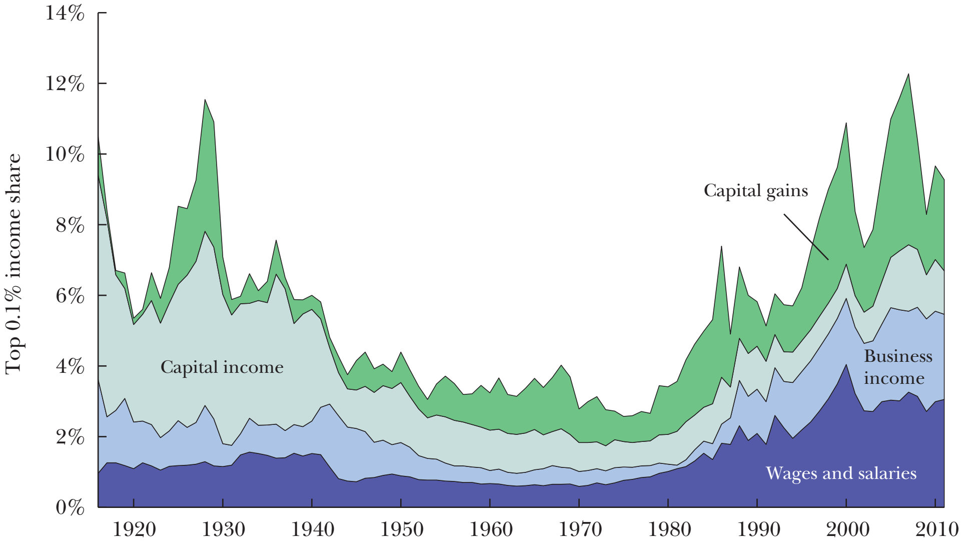
Two key results stand out. First, until recently, there is remarkably little growth in the average GDP per person at the top: the value in 1913 is actually *higher* than the value in 1977. Instead, all the growth until around 1960 occurs in the bottom 99.9 percent. Second, this pattern changed in recent decades. For example, average growth in GDP per person for the bottom 99.9 percent declined by around half a percentage point, from 2.3 percent between 1950 and 1980 to only 1.8 percent between 1980 and 2007. In contrast, after being virtually absent for 50 years, growth at the top accelerated sharply: GDP per person for the top 0.1 percent exhibited growth more akin to China's economy, averaging 6.86 percent since 1980. Changes like this clearly have the potential to matter for economic welfare and merit the attention they've received.

Labor Income Inequality

Basic Facts

One of the key papers documenting the rise in top income inequality is Piketty and Saez (2003), and it is appropriate to start with an updated graph from

Figure 2
The Top 0.1 Percent Income Share and Its Composition, 1916–2011



Source: These data are taken from the "data-Fig4B" tab of the September 2013 update of the spreadsheet appendix to Piketty and Saez (2003).

their paper. Figure 2 shows the share of income going to the top 0.1 percent of families in the United States, along with the composition of this income. Piketty and Saez emphasize three key facts seen in this figure. First, top income inequality follows a U-shaped pattern in the long term: high prior to the Great Depression, low and relatively steady between World War II and the mid-1970s, and rising since then, ultimately reaching levels today similar to the high levels of top income inequality experienced in the 1910s and 1920s. Second, much of the decline in top inequality in the first half of the 20th century was associated with capital income. Third, much of the rise in top inequality during the last several decades is associated with labor income, particularly if one includes "business income" in this category.

Theory

The next section of the paper will discuss wealth and capital income inequality. Here, motivated by the facts just discussed for the period since 1970, I'd like to focus on labor income inequality. In particular, what are the economic determinants of top labor income inequality, and why might they change over time and differ across countries?

At least since Pareto (1896) first discussed income heterogeneity in the context of his eponymous distribution, it has been appreciated that incomes at the top are well characterized by a power law. That is, apart from a proportionality factor to normalize units,

$$\Pr[\mathit{Income} > y] = y^{-1/\eta},$$

which means the fraction of people with incomes greater than some cutoff is proportional to the cutoff raised to some power. This is the defining characteristic of a Pareto distribution.

We can easily connect this distribution to the Piketty and Saez (2003) "top share" numbers. In particular, for the Pareto distribution just given, the fraction of income going to the top p percentiles equals $(100/p)^{\eta-1}$. In other words, the top share varies directly with the key exponent of the Pareto distribution, η. With $\eta = 1/2$, the share of income going to the top 1 percent is $100^{-1/2} = .10$, or 10 percent, while if $\eta = 2/3$, this share is $100^{-1/3} \approx 0.22$, or 22 percent. An increase in η leads to a rise in top income shares. Hence this parameter is naturally called a measure of Pareto inequality. In the US economy today, η is approximately 0.6.

A theory of top income inequality, then, needs to explain two things: (i) why do top incomes obey a Pareto distribution, and (ii) what economic forces determine η? The economics literature in recent years includes a number of papers that ask related questions. For example, Gabaix (1999) studies the so-called Zipf's Law for city populations: why does the population of cities follow a Pareto distribution, and why is the inequality parameter very close to 1? Luttmer (2007) asks the analogous question for firms: why is the distribution of employment in US firms a Pareto distribution with an inequality parameter very close to 1? Here, the questions are slightly different: Why might the distribution of income be well represented by a Pareto distribution, and why does the inequality parameter change over time and differ across countries? Interestingly, it turns out that there is a lot more inequality among city populations or firm employment than there is among incomes (their η's are close to 1 instead of 0.6). Also, the size distribution of cities and firms is surprisingly stable when compared to the sharp rise in US top income inequality.

From this recent economics literature as well as from an earlier literature on which it builds, we learn that the basic mechanism for generating a Pareto distribution is surprisingly simple: *exponential growth that occurs for an exponentially distributed amount of time leads to a Pareto distribution.*[2]

To see how this works, we first require some heterogeneity. Suppose people are exponentially distributed across some variable x, which could denote age or experience or talent. For example, $\Pr[\mathit{Age} > x] = e^{-\delta x}$, where δ denotes the death rate in the population. Next, we need to explain how income varies with age in the population. A natural assumption is exponential growth: suppose income rises exponentially with age (or experience or talent) at rate μ, that is, $\mathit{Income} = e^{\mu x}$. In

[2] Excellent introductions to Pareto models can be found in Mitzenmacher (2003), Gabaix (2009), Benhabib (2014), and Moll (2012b). Benhabib traces the history of Pareto-generating mechanisms and attributes the earliest instance of a simple model like that outlined here to Cantelli (1921).

this case, the log of income is just proportional to age, so the log of income obeys an exponential distribution with parameter δ/μ.

Next, we use an interesting property: if the log of income is exponential, then the level of income obeys a Pareto distribution:[3]

$$\Pr[\mathit{Income} > y] = y^{-\delta/\mu}.$$

Recall from our earlier discussion that the Pareto inequality measure is just the inverse of the exponent in this equation, which gives

$$\eta_{\text{income}} = \mu/\delta.$$

The Pareto exponent is increasing with μ, the rate at which incomes grow with age, and decreasing in the death rate δ. Intuitively, the lower is the death rate, the longer some lucky people in the economy can benefit from exponential growth, which widens Pareto inequality. Similarly, faster exponential growth across ages (which might be interpreted as a higher return to experience) also widens inequality.

This simple framework can be embedded in a richer model to produce a theory of top income inequality. For example, in Jones and Kim (2014) we build a model along these lines in which both μ and δ are endogenous variables that respond to changes in economic policy or technology. In our setup, x corresponds to the human capital of entrepreneurs. Entrepreneurs who put forth more effort cause their incomes to grow more rapidly, corresponding to a higher μ. The death rate δ is an endogenous rate of creative destruction by which one entrepreneur is displaced by another. Technological changes that make a given amount of entrepreneurial effort more effective, such as information technology or the worldwide web, will increase top income inequality. Conversely, exposing formerly closed domestic markets to international competition may increase creative destruction and reduce top income inequality. Finally, the model also incorporates an important additional role for luck: the richest people are those who not only avoid the destruction shock for long periods, but also those who benefit from the best idiosyncratic shocks to their incomes. Both effort and luck play central roles at the top, and models along these lines combined with data on the stochastic income process of top earners can allow us to quantify their comparative importance.

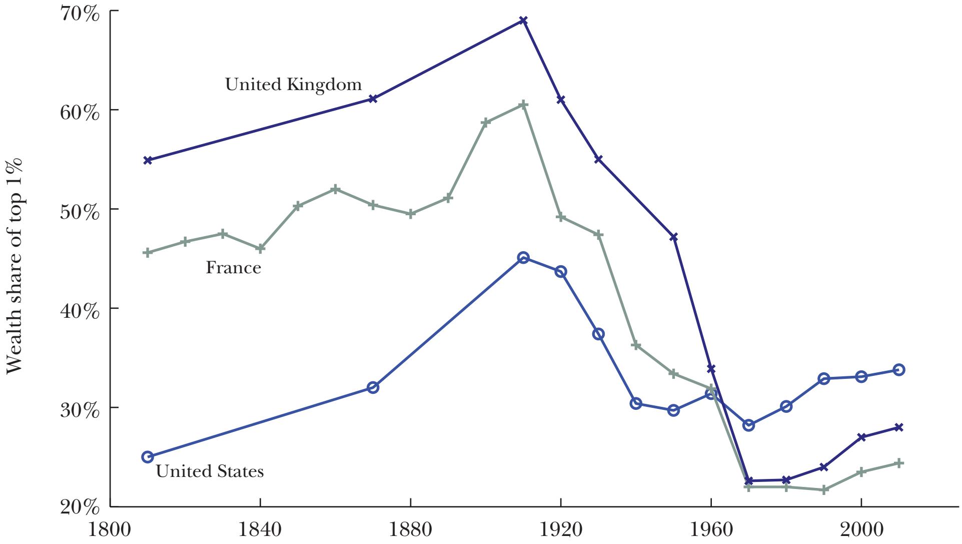
Wealth Inequality

Basic Facts

Up until this point, we've focused on inequality in labor income. Piketty's (2014) book, in contrast, is primarily about wealth, which turns out to be a more

[3] This derivation is explained in more detail in the online Appendix. Jones (2014) is available at http://www.stanford.edu/~chadj/SimpleParetoJEP.pdf and at the journal's website, http://e-jep.org.

Figure 3
Wealth Shares of the Top 1% in Three Countries, 1800 to 2010



Source: Supplementary Table S10.1 for chapter 10 of Piketty (2014), available at: http://piketty.pse.ens.fr/capital21c.
Note: The figure shows the share of aggregate wealth held by the richest 1 percent of the population.

difficult subject. Models of wealth are conceptually more complicated because wealth accumulates gradually over time. In addition, data on wealth are more difficult to obtain. Income data are "readily" (in comparison only!) available from tax authorities, while wealth data are gathered less reliably. For example, common sources include estate taxation, which affects an individual infrequently, or surveys, in which wealthy people may be reluctant to share the details of their holdings. With extensive effort, Piketty assembles the wealth inequality data shown in Figure 3, and several findings stand out immediately.

First, wealth inequality is much greater than income inequality. Figure 3 shows that the top 1 percent of families possesses around 35 percent of wealth in the United States in 2010—a newer source (Saez and Zucman 2014) says 40 percent—versus around 17 percent of income. Put another way, the income cutoff for the top 1 percent is about $330,000—in the ballpark of the top salaries for academics. In contrast, according to the latest data from Saez and Zucman (2014), the wealth cutoff for the top 1 percent is an astonishing $4 million! Note that both groups include about 1.5 million families.

Second, wealth inequality in France and the United Kingdom is dramatically lower today than it was at any time between 1810 and 1960. The share of wealth held by the top 1 percent is around 25 or 30 percent today, versus peaks in 1910 of 60 percent or more. Two world wars, the Great Depression, the rise of progressive taxation—some combination of these and other events led to an astonishing drop in wealth inequality both there and in the United States between 1910 and 1965.

Third, Figure 3 shows that wealth inequality has increased during the last 50 years, although the increase seems small in comparison to the declines just discussed. An important caveat to this statement applies to the United States: the data shown are those used by Piketty in his book, but Saez and Zucman (2014) have recently assembled what they believe to be superior data in the United States, and these data show a rise to a 40 percent wealth share for the US top 1 percent by 2010 (as mentioned earlier), much closer to the earlier peak in the first part of the 20th century.

Theory

A substantial and growing body of economic theory seeks to understand the determinants of wealth inequality.[4] Pareto inequality in wealth readily emerges through the same mechanism we discussed in the context of income inequality: exponential growth that occurs over an exponentially distributed amount of time. In the case of wealth inequality, this exponential growth is fundamentally tied to the interest rate, r: in a standard asset accumulation equation, the return on wealth is a key determinant of the growth rate of an individual's wealth. On the other hand, this growth in an individual's wealth occurs against a backdrop of economic growth in the overall economy. To obtain a variable that will exhibit a stationary distribution, one must normalize an individual's wealth level by average wealth per person or income per person in the economy. If average wealth grows at rate g, which in standard models will equal the growth rate of income per person and capital per person, the normalized wealth of an individual then grows at rate $r - g$. This logic underlies the key $r - g$ term for wealth inequality that makes a frequent appearance in Piketty's book. Of course, r and g are potentially endogenous variables in general equilibrium so—as we will see—one must be careful in thinking about how they might vary independently.

To be more specific, imagine an economy of heterogeneous people. The details of the model we describe next are given in Jones (2014). But the logic is straightforward. To keep it simple, assume there is no labor income and that individuals consume a constant fraction α of their wealth. As discussed above, wealth earns a basic return r. However, wealth is also subject to a wealth tax: a fraction τ is paid to the government every period. With this setup, the individual's wealth grows exponentially at a constant rate $r - \tau - \alpha$. Next, assume that average wealth per person (or capital per person) grows exogenously at rate g, for example in the context of some macro growth model. The individual's normalized wealth then grows exponentially

[4] References include Wold and Whittle (1957), Stiglitz (1969), Huggett (1996), Quadrini (2000), Castaneda, Diaz-Gimenez, and Rios-Rull (2003), Benhabib and Bisin (2006), Cagetti and Di Nardi (2006), Nirei (2009), Benhabib, Bisin, and Zhu (2011), Moll (2012a), Piketty and Saez (2012), Aoki and Nirei (2013), Moll (2014), and Piketty and Zucman (2014).

at rate $r - g - \tau - \alpha > 0$. This is the basic "exponential growth" part of the requirement for a Pareto distribution.

Next, we obtain heterogeneity in the simplest possible fashion: assume that each person faces a constant probability of death, d, in each period. Because Piketty (2014) emphasizes the role played by changing rates of population growth, we'll also include population growth, assumed to occur at rate n. Each new person born in this economy inherits the same amount of wealth, and the aggregate inheritance is simply equal to the aggregate wealth of the people who die each period. It is straightforward to show that the steady-state distribution of this birth-death process is an exponential distribution, where the age distribution is $\Pr[Age > x] = e^{-(n+d)}x$. That is, the age distribution is governed by the birth rate, which equals $n + d$. The intuition behind this formulation is that a fraction $n + d$ of new people are added to the economy each instant.

We now have exponential growth occurring over an exponentially distributed amount of time. The model we presented in the context of the income distribution suggested that the Pareto inequality measure equals the ratio of the "growth rate" to the "exponential distribution parameter" and that logic also holds for this model of the wealth distribution. In particular, wealth has a steady-state distribution that is Pareto with

$$\eta_{\text{wealth}} = \frac{r - g - \tau - \alpha}{n + d}.$$

An equation like this is at the heart of many of Piketty's statements about wealth inequality, for example as measured by the share of wealth going to the top 1 percent. Other things equal, an increase in $r - g$ will increase wealth inequality: people who are lucky enough to live a long time—or are part of a long-lived dynasty—will accumulate greater stocks of wealth. Also, a higher wealth tax will lower wealth inequality. In richer frameworks that include stochastic returns to wealth, the super-rich are also those who benefit from a lucky run of good returns, and a higher variance of returns will increase wealth inequality.

Can this class of models explain why wealth inequality was so high historically in France and the United Kingdom relative to today? Or why wealth inequality was historically much higher in Europe than in the United States? Qualitatively, two of the key channels that Piketty emphasizes are at work in this framework: either a low growth rate of income per person, g, or a low rate of population growth, n—both of which applied in the 19th century—will lead to higher wealth inequality.

Piketty (2014, p. 232) summarizes the logic underlying models like this with characteristic clarity: "[I]n stagnant societies, wealth accumulated in the past takes on considerable importance." On the role of population growth, for example, Piketty notes that an increase means that inherited wealth gets divided up by more offspring, reducing inequality. Conversely, a decline in population growth will concentrate wealth. A related effect occurs when the economy's per capita growth rate rises. In this case, inherited wealth fades in value relative to new wealth generated by economic growth. Silicon Valley in recent decades is perhaps an example worth considering. Reflections of these stories can be seen in the factors that determine η for the distribution of wealth in the equation above.

General Equilibrium

Whether changes in the parameters of models in this genre can explain the large changes in wealth inequality that we see in the data is an open question. However, one cautionary note deserves mention: the comparative statics just provided ignore the important point that arguably all the parameters considered so far are endogenous. For example, changes in the economy's growth rate g or the rate of the wealth tax τ can be mirrored by changes in the interest rate itself, potentially leaving wealth inequality unchanged.[5] To take another example, the fraction of wealth that is consumed, α, will naturally depend on the rate of time preference and the death rate in the economy. Because the parameters that determine Pareto wealth inequality are interrelated, it is unwise to assume that the direction of changing any single parameter will have an unambiguous effect on the distribution of wealth. General equilibrium forces matter and can significantly alter the fundamental determinants of Pareto inequality.

As one example, if tax revenues are used to pay for government services that enter utility in an additively separable fashion, the formula for wealth inequality in this model reduces to:

$$\eta_{\text{wealth}} = \frac{n}{n + d}.$$

See Jones (2014) for the details.[6] Remarkably, in this formulation the distribution of wealth is invariant to wealth taxes. In addition, the effect of population growth on wealth can actually go in the opposite direction from what we've seen so far. The intuition for this result is interesting: while in partial equilibrium, the growth rate of normalized wealth is $r - g - \tau - \alpha$, in general equilibrium, the only source of heterogeneity in the model is population growth. Newborns in this economy inherit the wealth of the people who die. Because of population growth, there are more newborns than people who die, so newborns inherit less than the average amount of wealth per capita. This dilution of the inheritance via population growth is the key source of heterogeneity in the model, and this force ties the distribution of wealth across ages at a point in time to population growth. Perhaps a simpler way of making the point is this: if there were no population growth in the model, newborns

[5] This relationship can be derived from a standard Euler equation for consumption with log utility, which delivers the result that $r - g - \tau = \rho$, where ρ is the rate of time preference. With log utility, the substitution and income effects from a change in growth or taxes offset and change the interest rate one for one.

[6] There are two key reasons for this result. The first is the Euler equation point made earlier, that $r - g - \alpha$ will be pinned down by exogenous parameters. The second is that the substitution and income effect from taxes cancel each other out with log utility, so the tax rate does not matter. For these two reasons, the numerator of the Pareto inequality measure for wealth, $r - g - \tau - \alpha$, simplifies to just n.

would each inherit the per capita amount of wealth in the economy. The accumulation of wealth by individuals over time would correspond precisely to the growth in the per capita wealth that newborns inherit, and there would be no inequality in the model despite the fact that $r > g$!

More generally, other possible effects on the distribution of wealth need to be considered in a richer framework. Examples include bequests, social mobility, progressive taxation, transition dynamics, and the role of both macroeconomic and microeconomic shocks. The references cited earlier make progress on these fronts.

To conclude this section, two points are worth appreciating. First, in a way that is easy to overlook because of our general lack of familiarity with Pareto inequality, Piketty is right to highlight the link between $r - g$ and top wealth inequality. That connection has a firm basis in economic theory. On the other hand, as I've tried to show, the role of $r - g$, population growth, and taxes is more fragile than this partial equilibrium reasoning suggests. For example, it is not necessarily true that a slowdown in either per capita growth or population growth in the future will increase inequality. There are economic forces working in that direction in partial equilibrium. But from a general equilibrium standpoint, these effects can easily be washed out depending on the precise details of the model. Moreover, these research ideas are relatively new, and the empirical evidence needed to sort out such details is not yet available.

Between-Inequality: Capital versus Labor

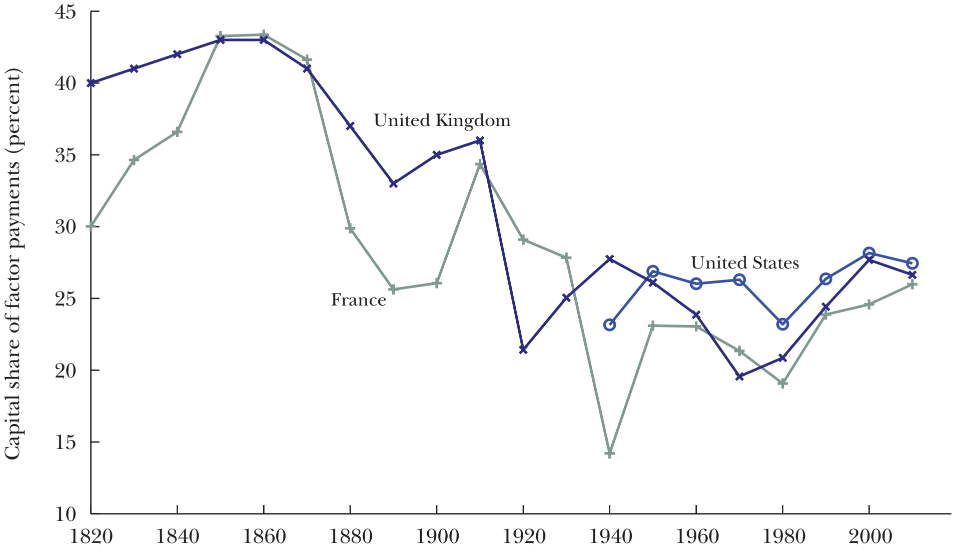
We next turn to between-inequality: how is income to capital versus income to labor changing, and how is the wealth–income ratio changing? This type of inequality takes on particular importance given our previous fact about within-inequality: most wealth is held by a small fraction of the population, which means that changes in the share of national income going to capital (that is, rK/Y) or in the aggregate capital–output ratio also contribute significantly to inequality. Whereas Pareto inequality describes how inequality at the top of the distribution is changing, this between-inequality is more about inequality between the top 10 percent of the population—who hold around 3/4 of the wealth in the United States according to Saez and Zucman (2014)—and the bottom 90 percent.

Basic Facts

At least since Kaldor (1961), a key stylized fact of macroeconomics has been the relative stability of factor payments to capital as a share of GDP. Figure 4 shows the long historical time series for France, the United Kingdom, and the United States that Piketty (2014) has assembled. A surprising point emerges immediately: prior to World War II, the capital share exhibits a substantial negative trend, falling from around 40 percent in the mid-1800s to below 30 percent. By comparison, the data since 1940 show some stability, though with a notable rise between 1980 and 2010. In Piketty's data, the labor share is simply one minus the capital share, so the

Figure 4

Capital Shares for Three Countries, 1820 to 2010



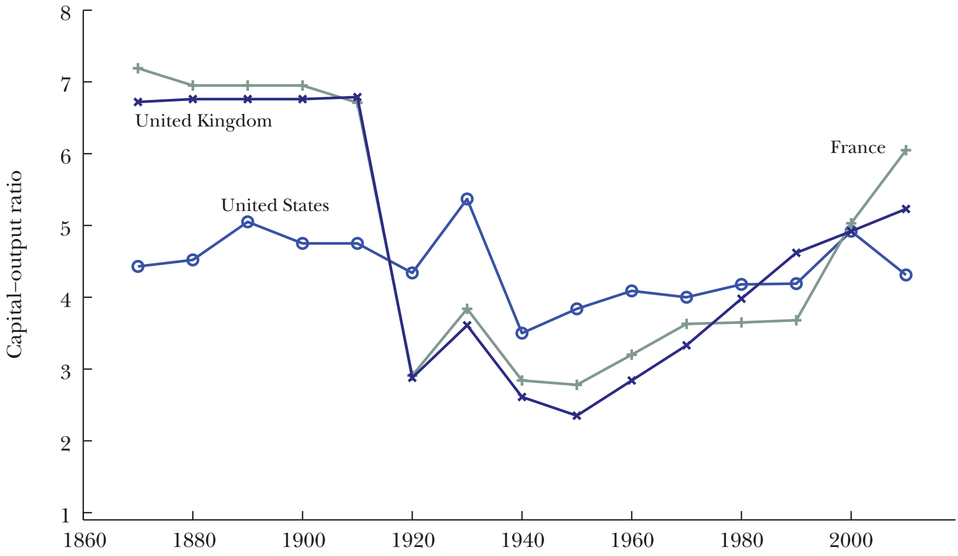
Source: For France and the United Kingdom, shares come from the supplementary tables for chapter 6 of Piketty (2014), available at http://piketty.pse.ens.fr/capital21c; for the United States, shares come from Piketty and Zucman (2014).
Note: Capital shares (including land rents) for each decade are averages over the preceding ten years.

corresponding changes in labor's share of factor payments can be read from this same graph.

Before delving too deeply into these numbers, it is worth appreciating another pattern documented by Piketty (2014). Figure 5 shows the capital–output ratio—the ratio of the economy's stock of machines, buildings, roads, land, and other forms of physical capital to the economy's gross domestic product—for this same group of countries, back to 1870. The movements are once again striking. France and the United Kingdom exhibit a very high capital–output ratio around 7 in the late 1800s. This ratio falls sharply and suddenly with World War I, to around 3, before rising steadily after World War II to around 6 today. The destruction associated with the two world wars and the subsequent transition dynamics as Europe recovers are an obvious interpretation of these facts. The capital–output ratio in the United States appears relatively stable in comparison, though still showing a decline during the Great Depression and a rise from 3.5 to 4.5 in the post–World War II period. These wonderful facts were not broadly known prior to Piketty's efforts.

Delving into the detailed data underlying these graphs, which Piketty (2014) generously and thoroughly provides, highlights an important feature of the data. By focusing on only two factors of production, capital and labor, Piketty includes land as a form of capital. Of course, the key difference between land and the rest of capital is that the quantity of land is fixed, while the quantity of other forms of

Figure 5
The Capital–Output Ratio



Source: Supplementary table S4.5 for chapter 4 of Piketty (2014), available at: http://piketty.pse.ens.fr/capital21c.

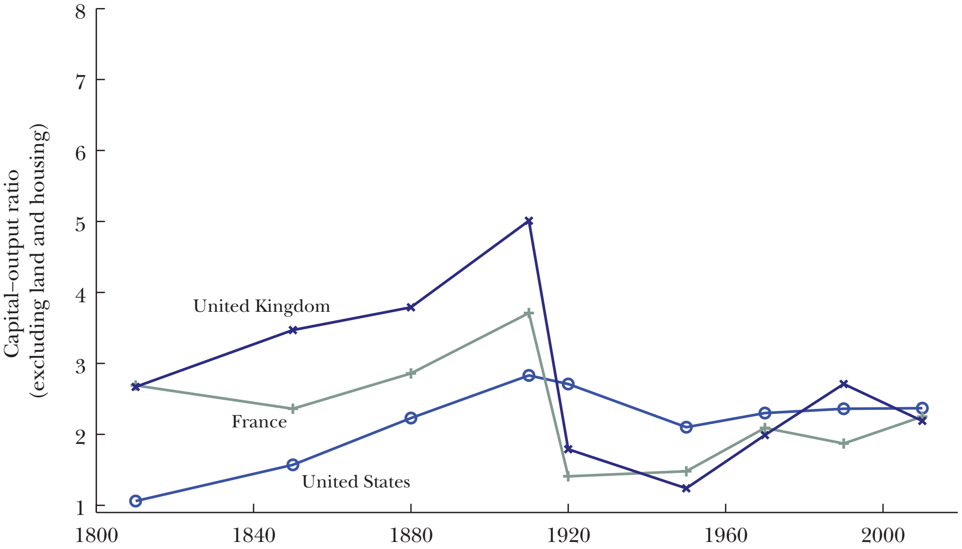
capital is not. For the purpose of understanding inequality between the top and the rest of the distribution, including land as a part of capital is eminently sensible. On the other hand, for connecting the data to macroeconomic theory, one must be careful.

For example, in the 18th and early 19th centuries, Piketty (2014) notes that rents paid to landlords averaged around 20 percent of national income. His capital income share for the United Kingdom before 1910 is taken from Allen (2007), with some adjustments, and shows a sharp decline in income from land rents (down to only 2 percent by 1910), which masks a rise in income from reproducible capital.

Similarly, much of the large swing in the European capital–output ratios shown in Figure 5 are due to land as well (in Piketty's book, Figures 3.1 and 3.2 make this clear). For example, in 1700 in France, the value of land equals almost 500 percent of national income versus only 12 percent by 2010. Moreover, the rise in the capital–output ratio since 1950 is to a great extent due to housing, which rises from 85 percent of national income in 1950 to 371 percent in 2010. Bonnet, Bono, Chapelle, and Wasmer (2014) document this point in great detail, going further to show that the rise in recent decades is primarily due to a rise in housing prices rather than to a rise in the quantity of housing.

As an alternative, consider what is called reproducible, nonresidential capital, that is the value of the capital stock excluding land and housing. This concept corresponds much more closely to what we think of when we model physical capital

Figure 6
The Capital–Output Ratio Excluding Land and Housing



Source: Supplementary tables S3.1, S3.2, and S4.2 for chapters 3 and 4 of Piketty (2014), available at: http://piketty.pse.ens.fr/capital21c.

in macro models. Data for this alternative are shown in Figure 6. In general, the movements in this measure of the capital–output ratio are more muted—especially during the second half of the 20th century. There is a recovery following the destruction of capital during World War II, but otherwise the ratio seems relatively stable in the latter period. In contrast, it is striking that the value in 2010 is actually lower than the value in several decades in the 19th century for both France and the United Kingdom. Similarly, the value in the United States is generally lower in 2010 than it was in the first three decades of the 20th century. I believe this is something of a new fact to macroeconomics—it strikes me as surprising and worthy of more careful consideration. I would have expected the capital-output ratio to be higher in the 20th century than in the 19th.

Stepping back from these discussions of the facts, an important point related to the "fundamental tendencies of capitalist economies," to use Piketty's language, needs to be appreciated. From the standpoint of overall wealth inequality, the declining role of land and the rising role of housing is not necessarily relevant. The inequality of wealth exists independent of the form in which the wealth is held. In the Pareto models of wealth inequality discussed in the preceding section, it turns out not to matter whether the asset that is accumulated is a claim on physical capital or a claim on a fixed aggregate quantity of land: the role of $r - g$ in determining the Pareto inequality measure η, for example, is the same in both setups. (The background models in Jones (2014) provide the details supporting this claim.) However, if one wishes to fit Piketty's long-run data to macroeconomic growth models—to say something about the shape of production functions—then it becomes crucial to distinguish between land and physical capital.

Theory

The macroeconomics of the capital-output ratio is arguably the best-known theory within all of macroeconomics, with its essential roots in the analysis of Solow (1956) and Swan (1956). The familiar formula for the steady-state capital–output ratio is $s/(n+g+\delta)$, where s is the (gross) investment share of GDP, n denotes population growth, g is the steady-state growth rate of income per person, and δ is the rate at which capital depreciates. Notice that this expression pertains to the ratio of *reproducible* capital—machines, buildings, and highways—and therefore is not strictly comparable to the graphs that Piketty (2014) reports, which include land.

In this framework, a higher rate of investment s will raise the steady-state capital–output ratio, while increases in population growth n, a rise in the growth rate of income per person g, or a rise in the capital depreciation rate δ would tend to reduce that steady-state ratio. Partly for expositional purposes, Piketty simplifies this formula to another that is mathematically equivalent: $\tilde{s}/\tilde{g}$, where $\tilde{g} = n + g$ and $\tilde{s}$ now denotes the investment rate net of depreciation, $\tilde{s} = s - \delta K/Y$. This more elegant equation is helpful for a general audience and gets the qualitative comparative statics right: in particular, Piketty emphasizes that a slowdown in growth—whether in per capita terms or in population growth—will raise the capital–output ratio in the long run. Piketty occasionally uses the simple formula to make *quantitative* statements: for example, if the growth rate falls in half, then the capital–output ratio will double (see Piketty's discussion beginning on p. 170). This statement is not correct and takes the simplification too far.[7]

It is plausible that some of the decline in the capital–output ratio in France and the United Kingdom since the late 1800s is due to a rise in the rate of population growth and the growth of income per person—that is, to a rise in $n+g$—and it is possible that a slowing growth rate of aggregate GDP in recent decades and in the future could contribute to a rise in the capital–output ratio. However, the quantitative magnitude of these effects is significantly mitigated by taking depreciation into account. These points, as well as a number of interesting related issues, are discussed in detail in Krusell and Smith (2014).

To see an example, consider a depreciation rate of 7 percent, a population growth rate of 1 percent, and a growth rate of income per person of 2 percent. In this case, in the extreme event that all growth disappears, the $n+g+\delta$ denominator of the Solow expression falls from 10 percent to 7 percent, so that the capital–output ratio increases by a factor of 10/7, or around 40 percent. That would be a large change, but it is nothing like the changes we see for France or the United Kingdom in Figure 5.

[7] In particular, it ignores the fact that $\tilde{s}$ will change when the growth rate changes, via the $\delta K/Y$ term.

One may also worry that these comparative statics hold the saving rate s constant. Fortunately, the case with optimizing saving is straightforward to analyze and gives similar results.[8] The bottom line from these examples is that qualitatively it is plausible that slowdowns in growth can increase the capital–output ratio in the economy, but the magnitudes of these effects should not be exaggerated.

The effect on between-inequality—that is, on the share of GDP paid as a return to capital—is even less clear. In the Cobb–Douglas example, of course, this share is constant. How then do we account for the empirical rise in capital's share since the 1980s? The research on this question is just beginning, and there are not yet clear answers. Recent papers studying the rise in the capital share in the last two decades include Karabarbounis and Neiman (2013), Elsby, Hobijn, and Şahin (2013), and Bridgman (2014).

Piketty himself offers one possibility, suggesting that the elasticity of substitution between capital and labor may be greater than one (as opposed to equaling one in the Cobb–Douglas case).[9] To understand this claim, look back at Figures 4 and 5. The fact that the capital share and the capital–output ratio move together, at least broadly over the long swing of history, is taken as suggestive evidence that the elasticity of substitution between capital and labor is greater than one. Given the importance of land in both of these time series, however, I would be hesitant to make too much of this correlation. The state-of-the-art in the literature on this elasticity is inconclusive, with some papers arguing for an elasticity greater than one but others arguing for less than one; for example, see Karabarbounis and Neiman (2013) and Oberfield and Raval (2014).

Conclusion

Through extensive data work, particularly with administrative tax records, Piketty and Saez and their coauthors have shifted our understanding of inequality in an important way. To a much greater extent than we've appreciated before, the dynamics of top income and wealth inequality are crucial. Future research combining this empirical evidence with models of top inequality is primed to shed light on this phenomenon.[10]

In *Capital in the Twenty-First Century*, Piketty suggests that the fundamental dynamics of capitalism will create a strong tendency toward greater inequality of wealth and even dynasties of wealth in the future unless this tendency is mitigated

[8] For example, with Cobb–Douglas production, $(r+\delta)K/Y=\alpha$, where α is the exponent on physical capital. With log utility, the Euler equation for consumption gives $r=\rho+g$. Therefore the steady state for the capital–output ratio is $\alpha/(\rho+g+\delta)$, which features similarly small movements in response to changes in per capita growth g.

[9] For example, see Piketty's (2014) discussion starting on p. 220.

[10] In this vein, it is worth noting that the Statistics of Income division of the Internal Revenue Service makes available random samples of detailed tax records in their public use microdata files, dating back to the 1960s (for more information on these data, see http://users.nber.org/~taxsim/gdb/).

by the enactment of policies like a wealth tax. This claim is inherently more speculative. Although the concentration of wealth has risen in recent decades, the causes are not entirely clear and include a decline in saving rates outside the top of the income distribution (as discussed by Saez and Zucman 2014), the rise in top labor income inequality, and a general rise in real estate prices. The theoretical analysis behind Piketty's prediction of rising wealth inequality often includes a key simplification in the relationships between variables: for example, assuming that changes in the growth rate g will not be mirrored by changes in the rate of return r, or that the saving rate net of depreciation won't change over time. If these theoretical simplifications do not hold—and there are reasons to be dubious—then the predictions of a rising concentration of wealth are mitigated. The future evolution of income and wealth, and whether they are more or less unequal, may turn on a broader array of factors.

I'm unsure about the extent to which $r - g$ will be viewed a decade or two from now as the key force driving top wealth inequality. However, I am certain that our understanding of inequality will have been enhanced enormously by the impetus—both in terms of data and in terms of theory—that Piketty and his coauthors have provided.

■ *I am grateful to Jess Benhabib, Xavier Gabaix, Jihee Kim, Pete Klenow, Ben Moll, and Chris Tonetti for helpful conversations and comments.*

References

Allen, Robert C. 2007. "Engel's Pause: A Pessimist's Guide to the British Industrial Revolution." Economics Series Working Papers 315, University of Oxford, Department of Economics, April.

Alvaredo, Facundo, Anthony B. Atkinson, Thomas Piketty, and Emmanuel Saez. N.d. The World Top Incomes Database. http://topincomes.g-mond.parisschoolofeconomics.eu/. Consulted October 1, 2013.

Aoki, Shuhei, and Makoto Nirei. 2013. "Pareto Distributions and the Evolution of Top Incomes in the U.S." MPRA Paper 47967, University Library of Munich, Germany.

Benhabib, Jess. 2014. "Wealth Distribution Overview." NYU teaching slides. http://www.econ.nyu.edu/user/benhabib/wealth%20distribution%20theories%20overview3.pdf.

Benhabib, Jess, and Alberto Bisin. 2006. "The Distribution of Wealth and Redistributive Policies." 2006 Meeting Papers 368, Society for Economic Dynamics.

Benhabib, Jess, Alberto Bisin, and Shenghao Zhu. 2011. "The Distribution of Wealth and Fiscal Policy in Economies with Finitely Lived Agents." *Econometrica* 79(1): 123–157.

Bonnet, Odran, Pierre-Henri Bono, Guillaume Chapelle, and Etienne Wasmer. 2014. "Does Housing Capital Contribute to Inequality? A Comment on Thomas Piketty's Capital in the 21st Century." Sciences Po Economics Discussion Paper 2014-07.

Bridgman, Benjamin. 2014. "Is Labor's Loss Capital's Gain? Gross versus Net Labor Shares." BEA Working Paper 0114.

Cagetti, Marco, and Mariacristina De Nardi. 2006. "Entrepreneurship, Frictions, and Wealth." *Journal of Political Economy* 114(5): 835–70.

Cantelli, F. P. 1921. "Sulle applicazioni del calcolo delle probabilita alla fisicamolecolare." *Metron* 1(3): 83–91.

Castaneda, Ana, Javier Diaz-Gimenez, and Jose-Victor Rios-Rull. 2003. "Accounting for the U.S. Earnings and Wealth Inequality." *Journal of Political Economy* 111(4): 818–57.

Elsby, Michael W. L., Bart Hobijn, and Ayşegül Şahin. 2013. "The Decline of the U.S. Labor Share." *Brookings Papers on Economic Activity*, no. 2, 1–63.

Gabaix, Xavier. 1999. "Zipf's Law for Cities: An Explanation." *Quarterly Journal of Economics* 114(3): 739–67.

Gabaix, Xavier. 2009. "Power Laws in Economics and Finance." *Annual Review of Economics* 1(1): 255–94.

Huggett, Mark. 1996. "Wealth Distribution in Life-Cycle Economies." *Journal of Monetary Economics* 38(3): 469–94.

Jones, Charles I. 2014. "Simple Models of Pareto Income and Wealth Inequality." http://www.stanford.edu/~chadj/SimpleParetoJEP.pdf.

Jones, Charles I., and Jihee Kim. 2014. "A Schumpeterian Model of Top Income Inequality." October 22. http://web.stanford.edu/~chadj/inequality.pdf.

Kaldor, Nicholas. 1961. "Capital Accumulation and Economic Growth." In *The Theory of Capital*, edited by F. A. Lutz and D. C. Hague, pp. 177–222. St. Martin's Press.

Karabarbounis, Loukas, and Brent Neiman. 2013. "The Global Decline of the Labor Share." *Quarterly Journal of Economics* 129(1): 61–103.

Krusell, Per, and Tony Smith. 2014. "Is Piketty's 'Second Law of Capitalism' Fundamental?" October 21. http://aida.wss.yale.edu/smith/piketty1.pdf.

Luttmer, Erzo G. J. 2007. "Selection, Growth, and the Size Distribution of Firms." *Quarterly Journal of Economics* 122(3): 1103–44.

Mitzenmacher, Michael. 2003. "A Brief History of Generative Models for Power Law and Lognormal Distributions." *Internet Mathematics* 1(2): 226–51.

Moll, Benjamin. 2012a. "Inequality and Financial Development: A Power-Law Kuznets Curve." August 12. http://www.princeton.edu/~moll/inequality.pdf.

Moll, Benjamin. 2012b. "Lecture 6: Income and Wealth Distribution." Teaching slides, ECO 521: Advanced Macroeconomics I, Princeton University. http://www.princeton.edu/~moll/ECO521Web/Lecture6_ECO521_web.pdf.

Moll, Benjamin. 2014. "Why Piketty Says $r - g$ Matters for Inequality." Teaching slides/lecture notes, Princeton University. http://www.princeton.edu/~moll/piketty_notes.pdf.

Nirei, Makoto. 2009. "Pareto Distributions in Economic Growth Models." IIR Working Paper 09-05, Institute of Innovation Research, Hitotsubashi University, July.

Oberfield, Ezra, and Devesh Raval. 2014. "MicroData and Macro Technology." August 29. https://sites.google.com/site/ezraoberfield/CESAggregation.pdf.

Pareto, Vilfredo. 1896. *Cours d'Économie Politique*. Geneva: Droz.

Piketty, Thomas. 2014. *Capital in the Twenty-first Century*. Harvard University Press.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in The United States, 1913–1998." *Quarterly Journal of Economics* 118(1): 1–41.

Piketty, Thomas, and Emmanuel Saez. 2012. "A Theory of Optimal Capital Taxation." NBER Working Papers 17989, April.

Piketty, Thomas, and Gabriel Zucman. 2014. "Capital is Back: Wealth-Income Ratios in Rich Countries, 1700–2010." *Quarterly Journal of Economics* 129(3): 1255–1310.

Quadrini, Vincenzo. 2000. "Entrepreneurship, Saving, and Social Mobility." *Review of Economic Dynamics* 3(1): 1–40.

Saez, Emmanuel, and Gabriel Zucman. 2014. "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data." NBER Working Paper 20625.

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

Stiglitz, Joseph E. 1969. "Distribution of Income and Wealth among Individuals." *Econometrica* 37(3): 382–97.

Swan, Trevor W. 1956. "Economic Growth and Capital Accumulation." *Economic Record* 32(2): 334–61.

Wold, Herman O. A., and Peter Whittle. 1957. "A Model Explaining the Pareto Distribution of Wealth." *Econometrica*, pp. 591–95.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 47–66

What Do We Know about the Evolution of Top Wealth Shares in the United States?†

Wojciech Kopczuk

In Piketty's (2014) prominent book, *Capital in the Twenty-First Century*, he argues that the concentration of wealth may become increasingly extreme in the future. As Piketty reminds us, the group of rentiers—people living off accumulated capital—has been historically large and politically and socially influential. Because so much of large fortunes end up being inherited, the current concentration of wealth is bound to predict at least weakly, and perhaps strongly, how important rentiers will be. Regardless of whether one buys into depictions such as "the rentier, enemy of democracy" (p. 422), the extent to which the well-off are going to rely on work versus rely on the returns to their wealth in the future is clearly important for assessing the extent to which a society will view itself as in some way a meritocracy.

Given that the US economy has experienced rising inequality in its income and earning distributions (for example, Piketty and Saez 2003; or see the symposium on "The Top 1 Percent" in the Summer 2013 issue of this journal), one would expect that the distribution of wealth would follow a similar path. However, available evidence on this topic is much more scant and conflicting than that on income and earnings. In fact, when Piketty (2014) reports direct estimates of wealth concentration for France, the United Kingdom, Sweden, and the United States in chapter 10 of his book, he finds as yet little evidence of dramatic increase in wealth concentration in any of these countries.

■ *Wojciech Kopczuk is Professor of Economics and of International and Public Affairs, Columbia University, New York City, New York. His email address is wojciech.kopczuk@columbia.edu.*

† To access the Data Appendix, visit
http://dx.doi.org/10.1257/jep.29.1.47

doi=10.1257/jep.29.1.47

In this paper, I discuss three different main methods for looking at the US wealth distribution: 1) the survey-based method using data from the Survey of Consumer Finances conducted by the Federal Reserve; 2) the estate multiplier method that uses data from estate tax returns to estimate wealth for the top of the wealth distribution; and 3) the capitalization method that uses information on capital income from individual income tax returns to estimate the underlying stock of wealth. At the time when Thomas Piketty wrote his book, only estimates based on the estate multiplier and the Survey of Consumer Finances were available; the capitalization method has been implemented by Saez and Zucman (2014) since the book was published. I also briefly comment on the usefulness of a fourth method: lists of high-wealth individuals, most notably the annual Forbes 400 list.

I will discuss the strengths and weaknesses of these approaches. I will focus in particular on a central difference in the estimates: the survey-based and estate tax methods suggest that the share of wealth held by the top 1 percent has not increased much in recent decades, while the capitalization method suggests that it has. I will offer some possible explanations for this divergence in findings: for example, questions over whether survey evidence on wealth captures those at the very top of the wealth distribution; varying estimates of the mortality rate of the very wealthy (which are necessary in projecting results from the estate tax to the broader population); sensitivity to rate-of-return assumptions; and changes in tax policy or business practices that would tend to alter the relationship between annual flows of income and accumulated stocks of wealth.

More broadly, as income inequality has grown in recent decades, the nature of wealth inequality has changed. Those in the top 1 percent of the US income and wealth distribution have less reliance on capital income and inherited wealth, and more reliance on income related to labor, than several decades ago. This transition can also help to explain why the methods of calculating wealth reach different results. These changes in the underlying sources and characteristics of high income and wealth must be the building blocks for understanding the connection between income and wealth inequality and whether, as predicted by Piketty (2014), the inequality of wealth and the importance of inherited wealth will dramatically rise in the future.

Basic Patterns in the Concentration of Wealth

There are four methods of measuring wealth at the very top of the distribution. First, one can carry out a survey that oversamples high-net-worth taxpayers. The Survey of Consumer Finances is the only source of that kind in the United States. Second, while the United States does not have an annual wealth tax (a few developed countries do—France and Norway in particular), it does have an estate tax. The estate tax records provide a snapshot of the distribution of wealth at the time of death. Third, while wealth itself is not reported to tax authorities, much of the capital income that wealth generates is taxable and observable, which provides an opportunity to estimate the underlying wealth distribution based on the annual flows of capital income. Finally, lists of named top wealth-holders exist—Forbes has published the best-known such list since 1982.

The coverage of these data sources varies in specific ways. In principle, the survey-based and capitalization methods allow for characterizing all (or, at least, most) of the wealth distribution. The estate tax approach is limited to drawing inferences based on the population subject to the tax. For most of the 20th century, this method allowed for constructing estimates for the top 1 percent, although changes since 2001 and especially since 2010 significantly reduced the coverage of the tax.[1] The lists of the wealthy are limited to the very small group of top wealth-holders and have nonsystematic coverage.

In terms of the time frames over which the data are available, estate tax and capitalization methods allow for constructing estimates going back to the beginning of the 20th century: the US income tax was introduced in 1913, and the estate tax was introduced in 1916. The Survey of Consumer Finances is available every three years starting with 1989, with precursor surveys available in 1962 (Survey of Financial Characteristics of Consumers) and 1983 (though it was also called the Survey of Consumer Finances, it had methodological differences relative to later surveys).[2] Differences in coverage and sampling suggest that 1962 and 1983 survey estimates should be treated with more caution than those for later years, especially for the top 1 percent. The capitalization series presented here is based on recent work of Saez and Zucman (2014) and covers the period from 1913–2012.

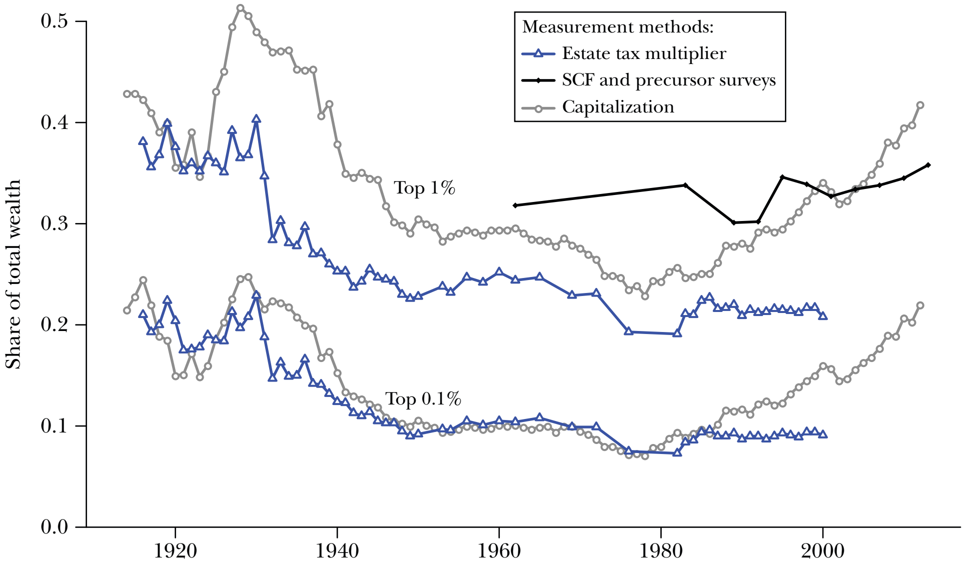
Each of the four methods has benefits and drawbacks that I will discuss in what follows. Before doing so, let us establish the basic facts. Figure 1 shows the evolution of the top 1 percent and top 0.1 percent of the wealth distribution using each of the methods that allow for constructing it. Figure 2 shows the evolution of the top 10 percent of the wealth distribution using the survey-based and capitalization methods and, separately, the wealth of the group from the 90th to 99th percentile—that is, the top 10 percent of the wealth distribution excluding the top 1 percent. Several observations are worth noting.

First, wealth is always highly concentrated. The share of wealth held by the top 10 percent has fluctuated between 65 and 85 percent of total wealth, the share of wealth held by the top 1 percent has ranged between 20 percent and as much as 45 percent of all wealth, and the share of wealth held by the top 0.1 percent ranged between less than 10 percent and as much as 25 percent.

[1] The estate tax series presented here is based on Kopczuk and Saez (2004a) and stops in 2000. Changes in the estate tax threshold reduced the coverage in subsequent years and will limit the applicability of this approach to groups significantly smaller than the top 1 percent.

[2] The series presented here were compiled by Roine and Waldenström (2015), and are in turn based on the work of Kennickell (2009b, 2011), Wolff (1996), and Lindert (2000). These estimates were extended to 2013 by Saez and Zucman (2014) following the Kennickell (2011) procedure. An unpublished paper by Scholz (2003) contains an alternative way of constructing wealth concentration estimates that generates very similar qualitative patterns. Related surveys are available for a few other years between 1962 and 1982, but they have not been used to estimate top wealth shares due to a small number of high-net-worth individuals.

Figure 1
Top 0.1% and Top 1% Wealth Shares

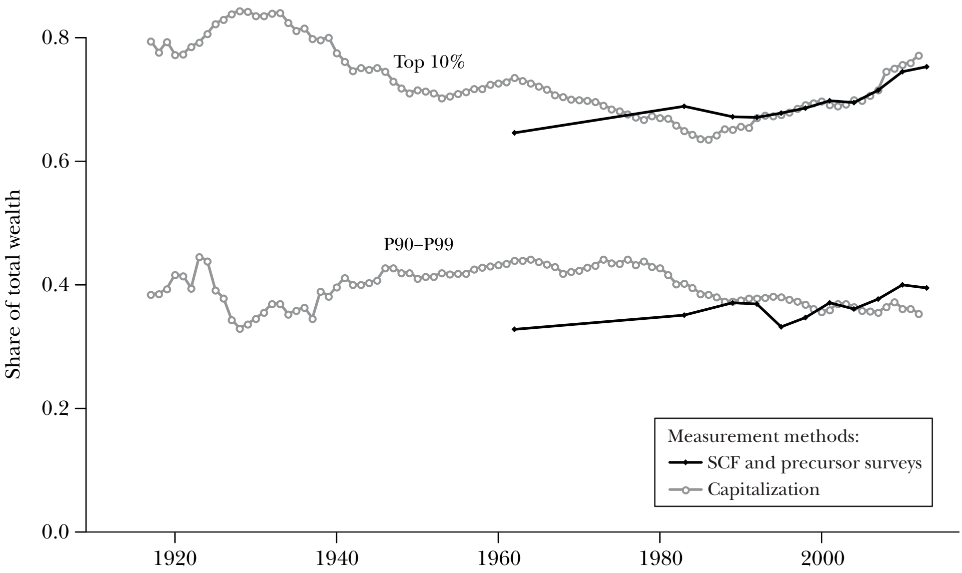


Source: Author using data described in the text.
Note: SCF is the Survey of Consumer Finances.

Second, the methods agree that the US wealth concentration peaked before the Great Depression and declined afterwards, staying relatively low at least until the 1980s. They do not necessarily agree on the timing though: the estate multiplier shows a rapid drop in the aftermath of the Great Depression, while the capitalization method shows more gradual adjustment, with rapid decline only in the late 1930s.

Third, the estate tax approach produces estimates that are lower than the other two approaches for the top 1 percent (estimates for the top 0.1 percent are much closer), but until the 1980s the two series available for that period move in a parallel fashion. There are conceptual differences that may generate different results from these approaches: for example, the estate tax multiplier method assigns wealth to individuals; the Survey of Consumer Finances (SCF) to households; and the capitalization method to "tax units." There are also differences in observability of assets. For example, tax evasion skews tax-based methods but not necessarily estimates from the SCF. Debt is observable on estate tax returns, but hard to capture by the capitalization method (debt is responsible for a reduction in the estate multiplier estimates of the top 1 percent share by more than 1 percentage point throughout and over 4 percentage points in the 1930s). Assets that do not generate taxable capital income have to be imputed in the capitalization approach.

Figure 2
Wealth Shares for Top 10% and 90th–99th Percentiles (P90–P99) in Terms of Wealth



Source: Author using data described in the text.
Note: SCF is the Survey of Consumer Finances.

Fourth, both the survey-based and the capitalization methods paint a very similar picture of the top 10 percent of the wealth distribution. Both indicate that the share of wealth held by the top 10 percent increased since the late 1980s.

Fifth, the different methods give diverging estimates since the 1980s, whether we look at the top 1 or top 0.1 percent of the wealth distribution. The methods that rely on direct measurement of wealth—that is, those based on the surveys and on the estate tax—show at best a small increase in the share of wealth held by the top 1 percent, while the capitalization methods shows a steep increase.

Sixth, given that the Survey of Consumer Finances and capitalization generate similar trends in recent years for the top 10 percent but different trends for the top 1 percent, it follows also that they do not coincide for the lower portion of the top 10 percent. The SCF shows a marked increase in the share of wealth going to P90–P99, while the capitalization method shows a decline.

These different approaches to estimating the distribution of wealth cover different periods of time and different parts of the distribution. They do not always paint the same picture, either. It is important then to understand the assumptions and the sources of data in order to understand weaknesses and strengths of different approaches. The next section discusses each of these four methods in more depth, and the following section then seeks to explain the discrepancies across the data series.

Four Methods of Measuring the Wealth Distribution

Survey of Consumer Finances

In a nutshell, the Survey of Consumer Finances is designed to measure household wealth. Bricker et al. (2014) and Kennickell (2009b, 2011) provide detailed overviews of its design. The definition of wealth in this survey includes all conventional categories of assets. Kennickell (2009b) concludes that the most important omissions are expected payments from defined benefit pension plans (naturally, Social Security wealth is also not accounted for), income streams from annuities or trusts, and human capital. In each case, these omissions are income-generating assets that are difficult or impossible to trade and that also escape the estate tax because they stop at death of the owner.

To cover the full wealth distribution in a way that accurately represents the concentration of wealth at the top end, the Survey of Consumer Finances supplements its random sample of the entire population with a stratified "list sample" derived from individual income tax returns. As a result, the survey significantly oversamples the very top of the wealth distribution. The sample, however, explicitly excludes individuals who belong to the Forbes 400 even if they are otherwise selected. Kennickell (2009a) notes that fewer than expected members of the Forbes 400 were selected and then disqualified, possibly because wealth in the Forbes sample may be held in trusts or by multiple family members, or because of errors in Forbes or issues with the Statistics of Income tax data that is relied on for stratification in the SCF.

A concern with the Survey of Consumer Finances is that the response rate among high-wealth individuals is only about 25 percent. Kennickell (2009a) discusses the response rate issue and the difficulties in reaching the very wealthy individuals, and concludes that the major difficulty in obtaining responses is the length of time that the interview takes. Given that this high-wealth sample is selected based on external income tax information, it is in principle possible to adjust for any potential nonresponse bias that varies systematically with observable characteristics: for example, if those, say, younger or with higher income were underrepresented because of a low response rate, those in these categories who did respond could be weighted more heavily. However, Kennickell (2009a) finds little evidence of nonresponse bias on observables.[3] In particular, he comments that refusal to fill out the survey (and various reasons for it) appears not related to the wealth index derived from income tax information that is relied on in sample design. Of course, one cannot eliminate the possibility that the sample is biased on some unobservable characteristics, but at least as the first pass, the sample does not appear biased in the dimensions that can be captured using income tax data.

[3] Verifying this point is the subject of ongoing work by the SCF staff, and I have confirmed that they still find that this conclusion holds in most recent surveys (personal communication).

Estate Tax Data

Since 1916—with the exception of 2010 when the estate tax requirement was eliminated for one year—estates of decedents with value exceeding a certain threshold are required to file an estate tax return. The threshold for the estate tax has varied significantly over time, but for most of the 20th century it corresponded to 1 percent or more of decedents being subject to the estate tax. In this way, the estate tax return provides a snapshot of wealth at the time of death for the population of sufficiently wealthy decedents.

A first practical difficulty in the estate tax approach is how to generalize from decedents to the full population. In Kopczuk and Saez (2004b), we provide extensive methodological discussion. The basic idea is to think of decedents as a sample from the living population. The individual-specific mortality rate m_i becomes the sampling rate. If m_i is known, the distribution for the living population can be simply estimated by reweighting the data for decedents by inverse sampling weights $1/m_i$, which are called "estate multipliers." Lampman (1962) was the first to provide such estimates for the US economy, although there are earlier estimates using UK data. In Kopczuk and Saez (2004a), we relied on confidential individual estate tax return data available at the IRS to construct such estimates for all years when they are available (1916–1945, a few years between 1946 and 1981, and 1982–2000) and supplemented it using data for a few other years between 1946 and 1981 for which detailed published tabulations exist.

The critical decision in applying the estate multiplier technique is the choice of mortality rates. While population mortality rates are relatively easy to observe by age and gender, mortality rates for the wealthy are known to be lower than those for the rest of the population, but are much harder to observe. In Kopczuk and Saez (2004a), we use estimated mortality differentials (by age and gender) between college-educated individuals (who are wealthier and longer living) and the full population at a single point in time (Brown, Liebman, and Pollet 2002) to adjust population mortality rates in all other years. The most worrisome feature of this approach is *not* that the mortality differentials for those with college education and for the wealthy are not the same: after all, as a first approximation such a difference would alter the level of the estimated wealth for the top groups, but would not necessarily affect the trend over time. A bigger concern is that the difference between mortality of college-educated and that of the wealthy may have changed over time. I will return to this issue when comparing capitalization and estate multiplier estimates.

Unlike the survey-based and capitalization methods, the estate tax method assigns wealth to individuals, rather than households. Depending on the composition of households (single vs. couple) across the distribution of wealth and on the division of assets within a household, this approach could in theory result in either higher or lower shares of top wealth percentiles relative to estimates based on a household distribution of wealth.

Another set of potential problems arises because the estate of a decedent may be different than wealth of an otherwise similar living person for various reasons.

As one example, an estate may have been diminished by a high level of end-of-life spending on health care. Estate tax data will reflect tax avoidance achieved by many high-wealth individuals through estate planning. The magnitude of the tax avoidance bias is difficult to assess, but some effect is clearly present; in Kopczuk (2013), I discuss available evidence. Certainly, there is a lot of estate tax planning and tax avoidance. At the same time, this phenomenon is not new, and there is no clear argument for why estate tax avoidance would have increased over time. Cooper (1979) dubbed the estate tax a "voluntary tax" in the 1970s, before any evidence would suggest that wealth inequality started growing. He showed that many aggressive estate tax planning techniques were possible at that time. Most of the loopholes he discussed can no longer be used, but new approaches have become available. The main constraint to aggressive tax planning, stressed by Schmalbeck (2001), is reluctance to relinquish control over wealth—effective estate tax planning inevitably corresponds to transfers with at least some irreversible aspects. Indeed, the available evidence suggests that there is too little tax planning in this context relative to what a fully tax-minimizing taxpayer would do (Kopczuk 2013).

Estate tax data that underlies the estate multiplier technique does not cover the full population. Hence, it cannot directly be used to provide an estimate of aggregate wealth, which in turn is necessary for constructing estimates of the share of wealth held by the top 0.1, top 1, or top 10 percent. In Kopczuk and Saez (2004a), we address this issue by constructing estimates of aggregate wealth using the Flow of Funds data. Saez and Zucman (2014) build on the same approach to construct aggregate wealth in their application of the capitalization method.

Capitalization Method

The idea behind the capitalization method of estimating wealth is straightforward. If we can observe capital income $k = rW$, where W is the underlying value of an asset and r is the known rate of return, then we can estimate wealth based on capital income and capitalization factor $1/r$ defined using the appropriate choice of rate of return. Many categories of capital income are subject to income taxation and hence income tax data may be used to implement this approach. Income tax data is "tax unit"–based; the unit may be a married couple or individual, with or without children, depending on tax-filing status selected by the taxpayer. Estimates obtained using this approach are likely closer to household (rather than individual) distribution of wealth. This method has a long history, although it has been rarely used in recent decades. Saez and Zucman (2014) implement and generalize this approach to construct what they refer to as "distributional Flow of Funds"—allocating aggregate wealth and its changes to different segments of the wealth distribution.

As one might expect, some practical difficulties arise in applying this approach. First, not all categories of assets generate capital income that appears on tax returns. For example, defined contribution pension plans do not generate taxable income as the funds accumulate. Owner-occupied housing does not generate annual taxable capital income, although it corresponds to property taxes that may be used to approximate its value in a rudimentary sense. The return on some types of investments is primarily taxed as capital gains if sold (capital gains are very problematic to deal with adequately, as discussed below) and are often held until death of the taxpayer, in which case they benefit from an increase in basis ("step up") and the underlying gain is never taxed on the individual level. Saez and Zucman (2014) report that capital income on tax returns represents only about one-third of the overall return to capital. The rest has to be imputed based on other information. Regarding capital gains, they either have to be explicitly accounted for, or capitalization factors need to be adjusted for pricing effects that correspond to unrealized returns. Works of art, closely-held businesses, and farm assets are examples of problematic categories with no easy fix. As a way of illustration, these categories account for 4 percent, 10 percent, and 3.7 percent of assets reported on estate tax returns in 2012 for taxpayers with over $20 million of assets (roughly a threshold for the top 0.1 percent of the wealth distribution). Also one needs to impute wealth in an explicit manner for categories of assets, such as personal residence, life-insurance, or pension funds, that do not generate income that is observable on individual tax returns. Saez and Zucman argue that these types of assets are not very important at the top of the distribution.

Second, both realized and expected returns to capital vary by asset, but only a very rough division of capital income is available on income tax returns: specifically, income tax returns include dividends, interest, capital gains, rents and royalties, and business income. Piketty (2014) argues that the rate of return to large portfolios exceeds the rate of return to smaller ones (see his discussion on pages 431 and 449, for example). Saez and Zucman (2014) effectively attribute such differences in rates of return to differences in portfolio composition between major assets classes corresponding to the few income streams that can be separately observed on tax returns, without allowing for correlation of rates of return within an asset class with the position in the income distribution.

Third, the capitalization approach assumes that capital income on tax returns on average represents normal return to wealth. There are a number of reasons for concern about this assumption, although it is hard to assess their importance. For example, some markets may be structured in favor of well-positioned individuals. An extreme example would be insider trading. A less-extreme example would be unequal access to high-yield investments, like those created by hedge funds that have high initial investment requirements. A benign but important example would be the extraordinary returns accruing to skilled entrepreneurs or investors. In each of these cases, the capitalization method would overestimate the level of wealth: instead of dividing the observed income by the actual realized rate of return, it would adjust it by a smaller, normal, rate of return.

Fourth, some types of income treated as return to capital on tax returns do not correspond to a person's underlying stock of wealth in a clear way. For example, the "carried interest" rule allows managers of certain investment funds to treat part of their compensation for managing assets as capital gains that are taxed at preferential rates. This is one of many examples of taxpayers acting on the strong incentive for those who face high marginal income tax rates to find ways to characterize their labor income as capital income. Other examples include payment through qualified stock options and certain choices about form of compensation in closely held firms. Such situations in which compensation is disguised as capital income are another reason why observed capital income might be higher than the normal rate would indicate, resulting in an overestimate of the underlying stock.

Fifth, wealthy individuals may in fact be those who received what, in retrospect, appears to be a very high rate of return. Obvious examples include successful technology companies—say Microsoft, Apple, or Google—that made their owners into billionaires. The capitalization method can capture the underlying stock of wealth after the valuation has already increased if assets pay, on average, normal dividends—although rapidly growing companies often do not pay dividends (Google still does not; Apple has only started in 2012; Microsoft initiated its dividend payouts in response to a dividend tax cut in 2003). But the capitalization method does not capture gain in the stock of equity wealth until individuals realize capital gains. Even if they do, such capital gains realized during explosive growth would correspond to extraordinary rates of return, but the capitalization method would interpret them as the outcome of a normal rate of return and hence would overestimate the underlying stock of wealth. It seems plausible that the prevalence of these types of issues is larger at the top of the distribution and that it has increased in recent decades with a rise in initial public offerings, weakening the attractiveness of the claim that such issues may somehow average out. Indeed, capital gains are an issue in general for the capitalization method, because income tax returns do not contain information about the holding period, which is necessary to capitalize them properly.

Sixth, the capitalization method is subject to biases due to tax avoidance. In fact, most tax avoidance/planning approaches that would skew estate tax data are going to leave a footprint in income tax data as well. As a trivial example, transfers of any income-generating assets would do so.

Despite these issues, the capitalization method produces estimates of wealth concentration that are parallel to the one obtained using the estate multiplier method until about 1986, as shown earlier in Figure 1. The key question, tackled in the next section, is to understand the source of differences in trends since then.

Saez and Zucman (2014) present a variety of validation checks for the capitalization method. For example, if one looks at the income reporting by foundations and applies this method, it does a good job of estimating the underlying wealth of the foundation. Of course, foundations are likely to be a poor counterfactual for the very wealthy individuals because foundations tend to be more diversified in their investments (in particular, for regulatory reasons) and they are nontaxable. Using matched income and estate data from the 1970s, Saez and Zucman also show that there is correspondence between wealth and capital incomes, which supports assumptions of the capitalization method. As another validation check, the Survey of Consumer Finances includes both income and wealth data, and the authors again show that the capitalization method allows the inference of wealth from the income data. Thus, there are surely reasons to be open to the possibility that the capitalization method may perform well in estimating wealth distribution.

Lists of the Wealthiest

Lists of the wealthiest Americans have the disadvantage of being based on valuations reported by journalists, which for a variety of reasons may contain errors or biases. However, one great advantage of such lists is that a researcher can identify specific people on the list and thus can identify whether their wealth comes from wages, other labor income, capital income, or inheritance. They also allow for looking at the age of top wealth-holders, their industry, and other factors.

The best-known of the lists of wealthy Americans is the Forbes 400. Using wealth as reported by Forbes, this group accounts for about a 2 percentage point increase in the total share of wealth at the top 1 percent (or the top 0.1 percent) between 1983 and 2013 (Saez and Zucman 2014). However, there are reasons to be concerned about the quality of this data. For example, Piketty (2014, pp. 441–443) is skeptical because he thinks that inherited wealth may be underrepresented. A direct comparison of estate tax returns and Forbes data by researchers from the IRS Statistics of Income Division (Johnson, Raub, and Newcomb 2013) finds that actual estates correspond to only about 50 percent of reported Forbes values. Part of this discrepancy may be due to tax avoidance and to a different way of allocating wealth (estate tax is individual, while Forbes often reports wealth for a "family"), but the gap is still very large. Possible reasons for overestimates in Forbes reports include difficulty in observing debt and differences in valuation approaches.

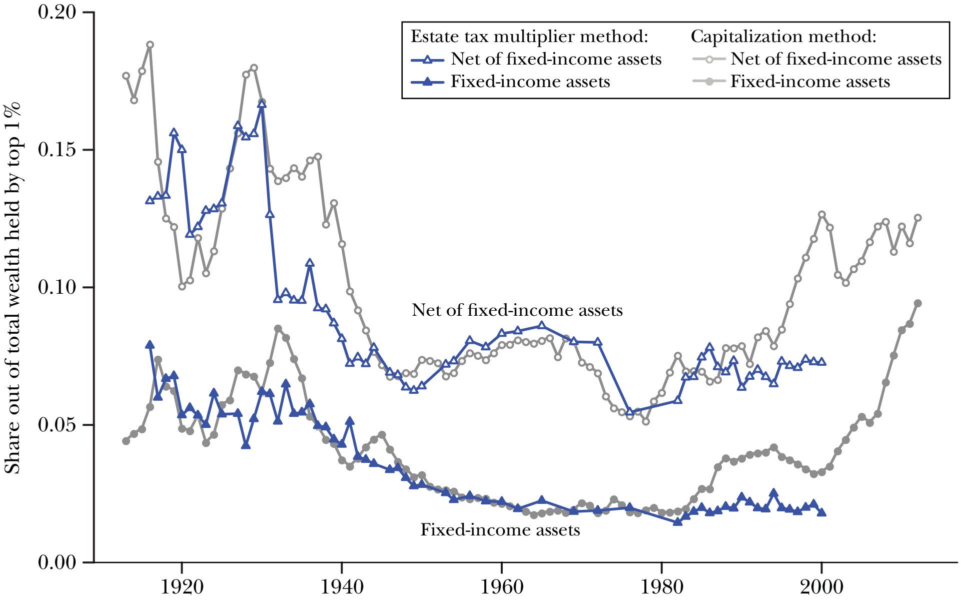
There are other historical lists going further back than Forbes. An impressive list of the 4,000 wealthiest Americans was published in 1892 by the *New York Tribune* newspaper. The website Classification of American Wealth (http://www.raken.com/american_wealth/) compiles many sources of information on top wealth-holders. Unfortunately, such sources are not systematic enough to allow for studying trends over time.

Understanding Discrepancies between Data Series

From about 1916 up until the 1960s, there are only two available approaches to estimating the evolving distribution of US wealth: the estate multiplier approach and the capitalization method. They agree that inequality in the distribution of wealth peaked in the 1920s, fell during the 1930s and into the 1940s, and then was mostly unchanged from the late 1940s up through the 1960s. As illustrated earlier in Figure 1, these data disagree on the level of wealth inequality during this time when looking at the top 1 percent, with the capitalization method usually providing higher estimates than the estate tax method. They are much closer for the smaller top 0.1 percent group. Possible straightforward explanations of the systematic difference in levels for the top 1 percent are differences in the unit of observation (individual versus "tax unit") and difficulty in observing debt on income tax returns.

There is one discrepancy during this time frame that is worth noting: the differing behaviors of the estate tax and capitalization series (as shown in Figure 1) around the time of the Great Depression. The estate tax approach shows an

Figure 3
Composition of the Top 0.1% Wealth Share



Source: Author using data described in the text.
Notes: The figure splits the estimated share of wealth accruing to the top 0.1 percent into two components: fixed income assets and everything else. The sum of the two components adds up to the share of the top 0.1 percent for the corresponding method. SCF is the Survey of Consumer Finances.

immediate decline in the share of wealth held by the top 1 percent during the Great Depression. Surprisingly, the capitalization method shows a smooth and fairly steady decline throughout the late 1920s through the 1940s, with the largest annual declines in the late 1930s and 1940s. This pattern resembles the Piketty and Saez (2003) finding that income inequality experienced the most rapid decline only in the 1940s.

Figure 3 shows what accounts for this difference. The figure splits the estimated share of wealth accruing to the top 0.1 percent into two components: fixed-income assets and everything else. Equities account for most of the latter category so that it primarily traces their dynamics; in particular, the share accounted for by real estate is fairly smooth and does not affect the qualitative pattern of the series. The sum of the two components adds up to the share of the top 0.1 percent for the corresponding method. Both methods show the decline in the non-fixed-income component (driven by equities) after 1929, although the decline in estate multiplier series is much steeper. Strikingly, the two series for the non-fixed-income component diverge throughout the 1930s. Furthermore, closer inspection of the underlying data available in online appendices to Saez and Zucman (2014) reveals that the capitalization factor for fixed income increases dramatically after 1929, reflecting lower yields, and that this effect is behind the temporary increase in the fixed-income component visible on Figure 3 in the early 1930s. The overall result is a relatively gentle decline in the overall share of the top 0.1 percent visible on Figure 1. Still, the *increase* in the share accounted for by the value of fixed-income assets in the capitalization series on Figure 3 nevertheless corresponds to about a 10 percent *decline* in the real value of such assets between 1930 and 1932.

There is of course the question of which series does a better job in representing dynamics over this period. Given similar dynamics of the two series before and after this episode and given that the estate tax captures wealth directly while the capitalization series relies on hard-to-verify assumptions about the relationship between capital income and underlying stock, it seems reasonable to suspect that the latter approach has trouble picking up distributional dynamics in the aftermath of Great Depression. In particular, it is hard to see why the estate tax series would have exaggerated the extent of decline in non-fixed-income assets between 1930 and 1932.

From about 1960 up through the early 1980s, some survey-based evidence on the wealth distribution becomes available through predecessors of the modern Survey on Consumer Finances. Together with the estimates from the estate tax approach and the capitalization method, the general pattern is that the level of inequality of the wealth distribution remains relatively unchanged throughout this period—although there is again a difference in the levels produced by the alternative methods as far as the top 1 percent is concerned (though the top 0.1 percent coincides remarkably well for capitalization and estate multiplier approaches).

However, for the period since about 1986, the trend in the distribution of wealth differs across these approaches. Estimates of the distribution of wealth based on the Survey of Consumer Finances and the estate tax method show little or no rise in the share of total wealth held by the top 1 percent in the last 30 years, while the capitalization approach finds a substantial rise (as shown earlier in Figure 1). In addition, the Survey of Consumer Finances data shows that the share of wealth received by the 90th to 99th percentile is rising in recent years, while the capitalization method suggests that the share of wealth for this group is falling.

How can these differences be explained? Some of the possible explanations include lower mortality rates for the wealthy (which could lead to biases in the estate tax method), concerns over survey representativeness (which could lead to biases in the survey-based method), trends in the bias in the rate of return assumptions under the capitalization method, and changes in the relationship between wealth and individual capital income on tax forms driven, for example, by changes in tax law (which could lead to biases in the capitalization method) or tax avoidance (which would affect both capitalization and estate multiplier approaches).

Composition of Top Wealth and Tax Incentives

A potential problem with the two tax-based approaches arises due to changes in tax incentives over the years. First, both approaches may be skewed by tax avoidance and evasion. While this would lead to understating the level of concentration, it is less clear that this would make a big difference for the trends because tax avoidance is hardly a new phenomenon and there is no clear presumption that it has secularly increased or declined over time. While international tax sheltering may be perhaps a bigger issue nowadays, corporate tax sheltering has likely been a much bigger issue in the past. The notion that tax avoidance has increased over time is also hard to reconcile with the evolution of tax rates. The top marginal income tax rate was above 60 percent from mid 1930s and 1981, and reached as high as 94 percent at its peak. It was then dramatically cut to 28 percent between 1981 and 1986 and remained below 40 percent ever since. Furthermore, tax avoidance is likely to affect both methods simultaneously. In particular, avoiding the estate tax usually entails transfer of assets and often income associated with them, so that it is likely to affect both estate multiplier and capitalization methods together.

Certain specific tax events appear important in understanding the discrepancy between the data series. The Tax Reform Act of 1986 in particular created an incentive to shift income from corporate to individual tax returns in a way that generated a massive behavioral response (Gordon and Slemrod 2000). The single largest short-term increase in top income shares according to Piketty and Saez (2003) takes place between 1986 and 1988 and reflects precisely this incentive. This is also the exact time when the capitalization measure of wealth begins to drift upward. There is no similar response at that point in time in estate multiplier estimates of wealth. This observation suggests the possibility that the capitalization method of estimating wealth, which is based on income-tax sources of information, may be responsive to tax-driven behavior in reporting or realization of capital income in ways that direct measures of wealth are not. More generally, changes in incentives and the repeal of the key provisions that had been behind some pre-1986s corporate tax shelters (such as the repeal of the "General Utilities doctrine") likely increased the extent to which wealth is revealed on individual-income (rather than corporate) tax data. These developments also potentially explain why the Survey of Consumer Finances—which, at least in principle, should not be biased by changes in tax treatment—yields larger wealth concentration in the top 1 percent in the 1960s and early 1980s than the capitalization method does, and why this difference disappears over time.

As in the aftermath of the Great Depression, the discrepancy between the two data series may also be traced to discrepancy in the composition of top wealth shares. As Figure 3 demonstrates, the sharp separation in the two series in 1986 is initially driven by the fixed-income component. Two incentives associated with the Tax Reform Act of 1986 may offer a potential explanation here. First, the reform significantly reduced deductibility of interest payments and may have increased net capital income reported on income tax returns, thereby driving up the estimate of its share under the capitalization method. Second, the shift from a corporate to an individual income tax base should have led to increases in all types of business-based income, including categories classified as fixed income.

Going forward, the estate tax series appears to completely miss the late 1990s stock market bubble and so does the Survey of Consumer Finances (although the infrequent timing of that survey may offer a partial explanation here), while the bubble is clearly visible in the capitalization series. This is very puzzling. It is possible that the estate tax somehow misses owners of successful tech companies who are relatively young and not likely to die, although in principle it should not be an issue since the observations for the few young individuals who do die would just end up being heavily weighted. One would also think that portfolios of other individuals would be partially invested in tech stocks, so that the run-up should be visible. None of these appears to be the case. One potential explanation is that estates may elect so called "alternate valuation" under which assets are valued at a later date than death (though, generally within a year)—this could result in smoothing the peak of the bubble, but it would be unlikely to eliminate its presence altogether. Hence, this piece of evidence appears to support the capitalization method. However, it also simultaneously casts doubt on one of its assumptions: in order for the, clearly very rich, estate taxpayers to miss the run-up in stock prices due the tech bubble, their estates had to be insufficiently diversified relative to what the capitalization method assumes. Put differently, this piece of evidence supports the idea that very high capital incomes on individual tax returns reflect extraordinary rather than normal returns.

The most striking feature of the estimates for the 2000s is a huge run-up of fixed income-generating wealth in the capitalization series. In fact, this run-up accounts for virtually *all* of the increase in the share of the top 0.1 percent between 2000 and 2012 and most of the increase since 2003. The underlying change in taxable capital income (reported by Saez and Zucman 2014, in their figure 3) is nowhere as dramatic. The share of fixed-income in overall capital income actually falls, as would be expected when yields fall. Instead, the (almost) tripling of the fixed income component on Figure 3 (from 3.3 percent of total wealth in 2000 to 9.5 percent in 2012) is driven by an increase in the underlying capitalization factor from 24 to 96.6. This is precisely what the method is intended to do: as yields have declined, the capitalization method should weight the remaining income much more heavily. This increase—if real—would correspond to enormous rebalancing of the underlying portfolios of the wealthy throughout the 2000s. An alternative possibility is simply that the capitalization factors are difficult to estimate during periods of very low rates of return, resulting in a systematic bias.

Mortality Rates for the Wealthy

As noted earlier, projecting from estate taxes to the general population requires using a mortality rate: the approach treats those who have died as a representative sample from the population. However, the wealthy have a lower mortality risk than the general population. Indeed, Saez and Zucman (2014) cite evidence suggesting that socioeconomic mortality differentials for broad demographic groups may have increased in recent decades. Furthermore, to shed a light on mortality changes at the very top of the wealth distribution, they use confidential IRS data, and they report that the mortality of those who are college-educated is a good approximation of mortality for the top 10 percent of the wealth distribution but that this proxy overestimates mortality rates higher in the wealth distribution. For example, their mortality rate evidence implies that mortality rates for 65 to 79 year-old males who are in the top 1 percent of the distribution are three quarters of the mortality rates of those in the top 10 percent. These are enormous differences in mortality rates: to put them in perspective, this estimated differential in mortality is bigger than that between the top 10 percent of the wealth distribution and the population average. Furthermore, they show that this discrepancy has increased since the 1970s and argue that the implied bias in estate multiplier might be able to explain the difference in trends between the estate tax method and the capitalization method.[4]

This explanation is conceptually plausible, but the estimated gap in mortality rates for the very wealthy is both very large and unexplored elsewhere in the literature, so the subject clearly requires further research. For example, an alternative possible explanation for their finding of such a large mortality advantage at the very top of the wealth distribution rests on the following observation: by construction, they report mortality rates for individuals with high capital income (which they interpret as high wealth); if high capital income represents active rather than passive returns (because it is a form of compensation for actively running or managing a business, for example), then individuals with high capital income are partially selected on health—it is being healthy that allows them to be active beyond retirement. On the flip side, individuals who are sickly may instead have an incentive to engage in tax planning and not realize capital income; in particular, there is a strong tax incentive not to realize capital gains until death in order to benefit from the step up of the basis of capital gains at death. As I will argue in what follows, it is likely that individuals at the top of the wealth distribution have become increasingly self-made, so that one might plausibly expect that this type of selection has become stronger over time. In addition, even such large increases in the mortality advantage of the very wealthy are still not large enough to explain the divergence between the capitalization and estate multiplier methods after the mid 1980s.[5]

Inclusion of Top Wealth-Holders?

As noted earlier, the Survey of Consumer Finances explicitly excludes those who appear on the Forbes 400. Saez and Zucman (2014) argue that one reason for the discrepancy between the SCF and the capitalization-based wealth estimates is that the SCF misses some of these top wealth-holders. However, remember that with more than 100 million households in the United States, the top 1 percent of the

[4] Their evidence indicates that mortality assumptions in the Kopczuk and Saez (2004a) study of the estate-tax-based measures of the wealth distribution are not far off for the 1970s, which is also the time when the capitalization method using merged estate and income tax data produces consistent results.

[5] Assuming a Pareto distribution with parameter a, a proportional increase in mortality differentials by a factor of $1 + x$ everywhere would result in an increase in the top share implied by the estate multiplier method by a factor of $(1 + x)^{1/a}$. Taking the value of $x = 0.3$ (an extremely large value, about the maximum adjustment suggested by Saez and Zucman, 2014, for any age group) and $a = 1.5$ (from Kopczuk and Saez, 2004a), it would yield an approximately 20 percent proportional adjustment in shares—in 2000, it amounts to about 4 percentage points correction for the top 1 percent share and about 2 percentage points for the top 0.1 percent, way short of the discrepancy between capitalization and estate multiplier methods that transpired between the 1980s and 2000s.

wealth distribution involves more than 1 million households. Even if the Forbes 400 list is capturing the very tip-top accurately—and as noted earlier, that assumption is dubious—the change in the top 400 can only account for about 2 percentage points of the 15 percent increase in the wealth share of the top 1 percent from 1983–2012 that the capitalization approach yields.

Going further down the distribution beyond the top 400 and into the rest of the top 1 percent of households in the wealth distribution, it is certainly possible that the Survey of Consumer Finances does miss individuals beyond the top 400 and does not correct for it by adjusting its weighting scheme, although Kennickell (2009a) finds no evidence of that. The sampling scheme in the SCF is based on income tax information, and hence it effectively identifies the top wealth-holders in a similar way as the capitalization method does. In neither case is wealth observed a priori, but wealthy individuals are sampled based on prediction of wealth from income. If this sampling approach fails to represent the wealthy population adequately in the Survey of Consumer Finances, the capitalization method will face similar problems. Similarly, just as the SCF does not include wealth from annuities or return to human capital, the capitalization method of estimating wealth is also likely to exclude this wealth; to the extent that income from these forms of wealth is taxable on individual tax returns, it would usually be taxable as labor income.

Hence, it is unclear why this type of bias would generate growing discrepancy between wealth estimates based on the Survey of Consumer Finances and the capitalization method. Furthermore, if the capitalization method produces accurate results and the SCF somehow misses the trend, one still would need to explain why the SCF provides an estimate of the wealth held by the top 1 percent which exceed the estimates of the capitalization approach in the 1980s but falls below the estimates of the capitalization approach in the 2000s (as visible in Figure 1).

Another issue with the capitalization method lies in its estimates of the share of wealth for the 90th to 99th percentile, shown in Figure 2. While one cannot completely rule out heavy trends in nonresponse bias in the Survey of Consumer Finances lower down in the wealth distribution, my prior is that this is not a likely explanation. Assuming that the SCF is representative of wealth in the 90th to 99th percentile group—which is much easier to measure accurately than the top 1 percent—then the capitalization method is actually getting steadily worse in measuring wealth in that group. One potential explanation here may have to do with an increasing importance of wealth held in the form of defined contribution pension plans, which are not observed in the income tax data and instead are imputed by the capitalization method. But of course, if imputations matter so much for the group from the 90th to 99th percentile, they may also matter elsewhere in the wealth distribution. One should also note that estimates of wealth not at the top of the distribution (such as the share of the bottom 90 or 99 percent) should be treated with caution: because many forms of wealth held lower in the distribution (pensions, housing) do not generate taxable income and require imputations, such estimates are effectively residuals obtained by subtracting estimates of the wealth at the top of the distribution from the overall wealth and hence contain little independent information.

Overall, the existing evidence on what happened to the concentration of wealth in the last few decades is not conclusive. My preference is to rely on the survey-based approach using the Survey of Consumer Finances and the estate-tax approach, primarily because of the strong assumptions and imputations needed to apply the capitalization methods in a way that gives consistent results over time. But this is a lively area of research, and the interpretation and implementation of all three of these approaches to estimating the concentration of wealth continues to evolve.

The Interplay of Income and Wealth Inequality

If, as the Survey of Consumer Finances and estate tax multiplier approaches say, the wealth share of the top 1 percent has *not* been rapidly trending upward, how can we reconcile this with the clear-cut evidence of growing *income* inequality? If, on the other hand, the capitalization method gets things right, is there an economic explanation for why the other two approaches seem to miss the growth in wealth concentration? I suspect that the difficulty here lies in the nature of changing inequality. Certainly, if the top 1 percent of incomes and the top 1 percent of wealth were the same people, growth in income shares would be expected to correspond to growth in top wealth shares.

However, the US distribution of income has not been stable in recent decades. There has been an increasing concentration of earnings over time, especially at the very top of the income distribution, as observed by Piketty and Saez (2003) and reiterated by many other authors. In addition, the nature of top incomes has changed since the 1920s—the last time when the share of income going to the top 1 percent was this high. In recent years, income at the top levels has been dominated by labor income; back in the 1920s, it was dominated by capital income (Piketty and Saez 2003). This change in the sources of income at the top suggests that the relationship between income inequality and wealth inequality has likely changed too.

The importance of inheritances as the source of wealth at top of the wealth distribution peaked in the 1970s and has declined since then, according to our analysis in Edlund and Kopczuk (2009). Our primary evidence is based on the gender composition of estate taxpayers and the observation that inherited wealth is much more equally distributed between sons and daughters than self-made wealth is. At the extreme tail of the wealth distribution, the trend has been toward observing more men, hence revealing the increased importance of self-made wealth. We also provide supportive evidence from a number of other sources, including the Forbes 400 list, that shows that the importance of inheritance among the richest Americans has declined since 1982 when the list was first published. Kaplan and Rauh (2013) provide a more comprehensive analysis of the Forbes 400 list and reach a similar conclusion. These observations suggest that the top of the wealth distribution is in flux. Individuals who are wealthy nowadays are less likely to come from wealth than in the past and more likely to have reached the top through earnings or entrepreneurial success.

Because wealth is an accumulated stock, not an annual flow, its distribution is bound to move more slowly than earnings distribution. The last 30 years have likely seen a transition in the upper parts of the wealth distribution, and this transition may still be taking place. Such a transition is consistent with a number of potential explanations I have given for why estimates of the trend in wealth concentration have been inconsistent in recent decades. For example, the increased importance of self-made, busy, active individuals among top wealth-holders is a plausible conjecture for why there could be a trend toward nonresponse bias among the wealthiest in the Survey of Consumer Finances and difficulties in observing them on estate tax returns. It is also a plausible reason for why large capital incomes may be increasingly reflecting work rather than underlying assets—which would then explain why there might be an observed trend in the mortality differential between people with high capital incomes (who are selected on being active) and everybody else. Without taking a stand on which of the preceding stories is most empirically important, these changes can plausibly reconcile the differences in methods of estimating the concentration of wealth, regardless of which one turns out to be closest to being right.

The central challenge for future work is to go beyond measuring income and wealth separately to try to understand how the joint distributions of income and wealth have been evolving over the last few decades—a period that certainly does not represent a steady state. Recognizing that the sources of income and wealth have been evolving for top income- and wealth-holders is a first step to improving our understanding of the trends and economic forces behind those patterns.

■ *I benefited from comments from David Autor, Marty Feldstein, Chang-Tai Hsieh, Ulrike Malmendier, Ben Marx, Jim Poterba, John Sabelhaus, Emmanuel Saez, Bernard Salanie, Timothy Taylor, and Scott Winship. Financial support from the Research Council of Norway grant #239225/H20 is gratefully acknowledged.*

References

Bricker, Jesse, Lisa J. Dettling, Alice Henriques, Joanne W. Hsu, Kevin B. Moore, John Sabelhaus, Jeffrey Thompson, and Richard A. Windle. 2014. "Changes in U.S. Family Finances from 2010 to 2013: Evidence from the Survey of Consumer Finances." *Federal Reserve Bulletin* 100(4): 1–40.

Brown, Jeffrey R., Jeffrey B. Liebman, and Joshua Pollet. 2002. "Appendix: Estimating Life Tables that Reflect Socioeconomic Differences in Mortality." In *The Distributional Aspects of Social Security and Social Security Reform*, edited by Martin Feldstein and Jeffrey B. Liebman, 447–57. University of Chicago Press.

Cooper, George. 1979. *A Voluntary Tax? New Perspectives on Sophisticated Tax Avoidance*. Studies of Government Finance. Brookings Institution.

Davies, James B., and Anthony F. Shorrocks. 2000. "The Distribution of Wealth." Chap. 11 in *Handbook of Income Distribution*, edited by

Anthony B. Atkinson and François Bourguignons. New York: Elsevier.

Edlund, Lena, and Wojciech Kopczuk. 2009. "Women, Wealth and Mobility." *American Economic Review* 99(1): 146–78.

Feenberg, Daniel R., and James M. Poterba. 1993. "Income Inequality and the Incomes of Very High-Income Taxpayers: Evidence from Tax Returns." In *Tax Policy and the Economy*, vol. 7, edited by James M. Poterba, 145–77. MIT Press.

Gordon, Roger H., and Joel Slemrod. 2000. "Are 'Real' Responses to Taxes Simply Income Shifting Between Corporate and Personal Tax Bases?" Chap. 8 in *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, edited by Joel Slemrod. New York: Harvard University Press.

Johnson, Barry, Brian Raub, and Joseph Newcomb. 2010. "A Comparison of Wealth Estimates for America's Wealthiest Decedents Using Tax Data and Data from The Forbes 400." In *National Tax Association Proceedings of 103rd Annual Conference on Taxation*, 128–135.

Kaplan, Steven N., and Joshua D. Rauh. 2013. "Family, Education, and Sources of Wealth among the Richest Americans, 1982–2012." *American Economic Review* 103(3): 158–62.

Kennickell, Arthur B. 2009a. "Getting to the Top: Reaching Wealthy Respondents in the SCF," Paper prepared for the 2009 Joint Statistical Meetings, Washington, DC, Federal Reserve Board.

Kennickell, Arthur B. 2009b. "Ponds and Streams: Wealth and Income in the U.S., 1989 to 2007." Finance and Economics Discussion Series 2009-13, Federal Reserve Board.

Kennickell, Arthur B. 2011. "Tossed and Turned: Wealth Dynamics of U.S. Households 2007–2009." Finance and Economics Discussion Series 2011-51, Federal Reserve Board.

Kopczuk, Wojciech. 2013. "Taxation of Intergenerational Transfers and Wealth." In *Handbook of Public Economics*, vol. 5, edited by Alan J. Auerbach, Raj Chetty, Martin S. Feldstein, and Emmanuel Saez, 329–90. Elsevier.

Kopczuk, Wojciech, and Emmanuel Saez. 2004a. "Top Wealth Shares in the United States, 1916–2000: Evidence from Estate Tax Returns." *National Tax Journal* 57(2, part 2): 445–88.

Kopczuk, Wojciech, and Emmanuel Saez. 2004b. "Top Wealth Shares in the United States, 1916–2000: Evidence from Estate Tax Returns." NBER Working Paper 10399.

Lampman, Robert J. 1962. *The Share of Top Wealth-Holders in National Wealth, 1922–56*. Princeton University Press.

Lindert, Peter H. 2000. "Three Centuries of Inequality in Britain and America." In *Handbook of Income Distribution*, edited by Anthony B. Atkinson and Francois Bourguignon, 167–216. Elsevier, North Holland.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Harvard University Press.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118(1): 1–41.

Roine, Jesper, and Daniel Waldenström. 2015. "Long-Run Trends in the Distribution of Income and Wealth." Chap. 7 in *Handbook of Income Distribution*, vol. 2, edited by Anthony B. Atkinson and Francois Bourguignon. North Holland.

Saez, Emmanuel, and Gabriel Zucman. 2014. "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data." NBER Working Paper 20625, October.

Schmalbeck, Richard. 2001. "Avoiding Federal Wealth Transfer Taxes." Chap. 3 in *Rethinking Estate and Gift Taxation*, edited by William G. Gale, James R. Hines Jr., and Joel Slemrod. Brookings Institution Press.

Scholz, John Karl. 2003. "Wealth Inequality and the Wealth of Cohorts." https://www.russellsage.org/sites/all/files/u4/Scholz.pdf.

Wolff, Edward N. 1996. "International Comparisons of Wealth Inequality." *Review of Income and Wealth* 42(4): 433–51.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 67–88

Putting Distribution Back at the Center of Economics: Reflections on *Capital in the Twenty-First Century*

Thomas Piketty

When a lengthy book is widely discussed in academic circles and the popular media, it is probably inevitable that the arguments of the book will be simplified in the telling and retelling. In the case of my book *Capital in the Twenty-First Century* (2014), a common simplification of the main theme is that because the rate of return on capital r exceeds the growth rate of the economy g, the inequality of wealth is destined to increase indefinitely over time. In my view, the magnitude of the gap between r and g is indeed one of the important forces that can explain historical magnitudes and variations in wealth inequality: in particular, it can explain why wealth inequality was so extreme and persistent in pretty much every society up until World War I (for discussion, see Chapter 10 of my book). That said, the way in which I perceive the relationship between $r > g$ and wealth inequality is often not well-captured in the discussion that has surrounded my book—even in discussions by research economists.

In this essay, I will return to some of the themes of my book and seek to clarify and refocus the discussion concerning those themes. For example, I do not view $r > g$ as the only or even the primary tool for considering changes in income and wealth in the 20th century, or for forecasting the path of income and wealth inequality in the 21st century. Institutional changes and political shocks—which can be viewed as largely endogenous to the inequality and development process itself—played a major role in the past, and will probably continue to do so in the future. In addition, I certainly do not believe that $r > g$ is a useful tool for the discussion of rising inequality of labor income: other mechanisms

■ *Thomas Piketty is Professor of Economics at the Paris School of Economics, Paris, France. His email address is piketty@psemail.eu.*

http://dx.doi.org/10.1257/jep.29.1.67 doi=10.1257/jep.29.1.67

and policies are much more relevant here, for example, the supply and demand of skills and education. One of my main conclusions is that there is substantial uncertainty about how far income and wealth inequality might rise in the 21st century and that we need more transparency and better information about income and wealth dynamics so that we can adapt our policies and institutions to a changing environment.

My book is primarily about the history of the distribution of income and wealth. Thanks to the cumulative efforts of several dozen scholars, we have been able to collect a relatively large historical database on the structure of national income and national wealth, and the evolution of income and wealth distributions, covering three centuries and over 20 countries. The first objective of my book was to present this body of historical evidence and to analyze the economic, social, and political processes that can account for the evolutions that we observe in the various countries since the Industrial Revolution. I stress from the beginning that we have too little historical data at our disposal to be able to draw definitive judgments. On the other hand, at least we have substantially more evidence than we used to have.

My book is probably best described as an analytical historical narrative based upon this new body of evidence. In this way, I hope I can contribute to placing the study of distribution and of the long-run back at the center of economic thinking. Many 19th century economists, including Thomas Malthus, David Ricardo, and Karl Marx, put the distribution question at the center of political economy. However, they had limited data at their disposal, and so their approach was mostly theoretical. In contrast, since the mid-20th century, a number of economists, most notably Simon Kuznets and Anthony Atkinson, have been developing the possibility of an approach that blends theory with more data-intensive and historical approaches. This historical data collection project on which my book is based follows directly in the tradition of the pioneering works by Kuznets (1953) and Atkinson and Harrison (1978).

In this essay, I will take up several themes from my book that have perhaps become attenuated or garbled in the ongoing discussions of the book, and will seek to re-explain and re-frame these themes. First, I stress the key role played in my book by the interaction between beliefs systems, institutions, and the dynamics of inequality. Second, I briefly describe my multidimensional approach to the history of capital and inequality. Third, I review the relationship and differing causes between wealth inequality and income inequality. Fourth, I turn to the specific role of $r > g$ in the dynamics of wealth inequality: specifically, a larger $r - g$ gap will amplify the steady-state inequality of a wealth distribution that arises out of a given mixture of shocks. Fifth, I consider some of the scenarios that affect how $r - g$ might evolve in the 21st century, including rising international tax competition, a growth slowdown, and differential access by the wealthy to higher returns on capital. Finally, I seek to clarify what is distinctive in my historical and political economy approach to institutions and inequality dynamics, and the complementarity with other approaches.

Beliefs Systems, Institutions, and the Dynamics of Inequality

In my book, I attempt to study not only the dynamics of income and wealth inequality, but also the evolution of collective representations of social inequality in public discussions and political debates, as well as in literature and movies. I believe that the analysis of representations and beliefs systems about income and wealth is an integral and indispensable part of the study of income and wealth dynamics.

Indeed, a main conclusion of my analytical historical narrative is stated in the introduction of the book (p. 20, 35), that "one should be wary of any economic determinism in regard to inequalities of wealth and income . . . The history of the distribution of wealth has always been deeply political, and it cannot be reduced to purely economic mechanisms. . . . It is shaped by the way economic, social, and political actors view what is just and what is not, as well as by the relative power of those actors and the collective choices that result. It is the joint product of all relevant actors combined. . . . How this history plays out depends on how societies view inequalities and what kinds of policies and institutions they adopt to measure and transform them." As I wrote in a follow-up essay with a co-author: "In a sense, both Marx and Kuznets were wrong. There are powerful forces pushing alternatively in the direction of rising or shrinking inequality. Which one dominates depends on the institutions and policies that societies choose to adopt" (Piketty and Saez 2014, p. 842–43).

The role of political shocks and changing representations of the economy is especially obvious when one studies inequality dynamics during the 20th century. In particular (p. 20), "the reduction of inequality that took place in most developed countries between 1910 and 1950 was above all a consequence of war and revolution and of policies adopted to cope with these shocks. Similarly, the resurgence of inequality after 1980 is due largely to the opposite political shifts of the past several decades, especially in regard to taxation and finance."

I also try to show that belief systems about the distribution of income and wealth matter a great deal if one wants to understand the structure of inequality in the 18th and 19th centuries, and indeed in any society. Each country has its own intimate history with inequality, and I attempt to show that national identities play an important role in the two-way interaction between inequality dynamics and the evolution of perceptions, institutions, and policies.

I continually refer to a large number of other institutions and public policies that play a substantial role in my historical account of inequality dynamics across three centuries and over 20 countries. I emphasize the importance of educational institutions (in particular the extent of equal access to high-quality schools and universities) and of fiscal institutions (especially the chaotic advent of progressive taxation of income, inheritance, and wealth). Other examples of important factors include: the development of the modern welfare state; monetary regimes, central banking, and inflation; labor market rules, minimum wages, and collective bargaining; forced labor (slavery); colonialism, wars, and revolutions; expropriations, physical destruction, and privatizations; corporate governance and stakeholder rights; rent and other price controls (such as the prohibition or limitation of usury); financial deregulation and capital flows; trade policies; family transmission rules and legal property regimes; fertility policies; and many others.

A Multidimensional History of Capital and Inequality

A central reason that my book is relatively long is that I try to offer a relatively detailed, multidimensional history of capital and its metamorphosis. Capital ownership takes many different historical forms, and each of them involves different forms of institutions, rules, and power relations, which must be analyzed as such.

Theoretical models, abstract concepts, and equations (such as $r > g$, to which I return in greater detail below) also play a certain role in my analysis. However this role is relatively modest—as I believe the role of theory should generally be in the social sciences—and it should certainly not be exaggerated. Models can contribute to clarifying logical relationships between particular assumptions and conclusions but only by oversimplifying the real world to an extreme point. Models can play a useful role but only if one does not overestimate the meaning of this kind of abstract operation. All economic concepts, irrespective of how "scientific" they pretend to be, are intellectual constructions that are socially and historically determined, and which are often used to promote certain views, values, or interests. Models are a language that can be useful only if solicited together with other forms of expressions, while recognizing that we are all part of the same conflict-filled, deliberative process.

In particular, the notion of an aggregate capital stock K and of an aggregate production function $Y = F(K, L)$ are highly abstract concepts. From time to time, I refer to them. But I certainly do not believe that such grossly oversimplified concepts can provide an adequate description of the production structure and the state of property and social relations for any society. For example, I explain in Chapter 1, when I define capital and wealth (p. 47):

> Capital is not an immutable concept: it reflects the state of development and prevailing social relations of each society. . . . The boundary between what private individuals can and cannot own has evolved considerably over time and around the world, as the extreme case of slavery indicates. The same is true of property in the atmosphere, the sea, mountains, historical monuments, and knowledge. Certain private interests would like to own these things, and sometimes they justify this desire on grounds of efficiency rather than mere self-interest. But there is no guarantee that this desire coincides with the general interest.

More generally, I analyze the diversity of the forms taken by capital assets and the problems raised by property relations and market valorizations throughout history. I study in some length the many transformations in the nature of capital assets, from agricultural land to modern real estate and business and financial capital. Each type of asset has its own particular economic and political history and gives rise to different bargaining processes, power struggles, economic innovations, and social compromises.

For example, the fact that capital ownership and property rights are historically determined is particularly clear when I study the role of slave capital in the Southern United States before 1865, which can be viewed as the most extreme form of ownership and domination of owners over others (Chapter 4). A similar theme also becomes evident when I examine the lower stock market capitalization of German companies relative to their Anglo-American counterparts, a phenomenon that is certainly related to the fact that German shareholders need to share power with other stakeholders (workers, governments, nongovernment organizations, and others) somewhat more than in other countries (Chapter 5). This power-sharing apparently is not detrimental to the productive efficiency and exporting performance of German firms, which illustrates the fact that the market and social values of capital can often differ.

Other examples involve real estate capital and natural resource wealth—like oil. Large upward or downward movements of real estate prices play an important role in the evolution of aggregate capital values during recent decades, as they did during the first half of the 20th centuries (in particular, Chapters 3–6). This can in turn be accounted for by a complex mixture of institutional and technological forces, including rent control policies and other rules regulating relations between owners and tenants, the transformation of economic geography, and the changing speed of technical progress in the transportation and construction industries relative to other sectors. The issue of oil capital and its world distribution is rooted in the power relations and military protections that go with it (in particular in the Middle East), which also have consequences for the financial investment strategies followed by the corresponding sovereign wealth funds (discussed in Chapter 12).

The institutional analysis of property relations and capital assets also has international and public-sector dimensions. The hypertrophy of gross financial asset positions between countries, which is one of the main characteristics of the financial globalization process of recent decades, is a recurring theme of the book (Chapters 1–5, 12, 15, and 16). I analyze the very large magnitude of the net foreign assets positions reached by Britain and France at the height of their colonial empires, and I compare them to today's net positions of China, Japan, or Germany. I repeatedly stress that international property relations—the fact that economic actors in some countries own significant claims on real and financial assets in other countries—can be particularly complicated to regulate in a peaceful manner. This was certainly true during the colonization and decolonization periods. Issues of international property relations could erupt again in the future. The difficulty in dealing with extreme internal and external inequality certainly contributes to explaining the high political instability that has long plagued the development process in Latin American and African countries.

Public capital—which depends on the changing patterns and complex political histories of public investment and deficit trajectories and nationalization and privatization policies—also plays a critical role in the book (especially Chapters 3 and 4). I emphasize the sharp dissimilarities in country experiences (contrasting in particular the cases of Britain and France in the 18th and 19th centuries), as well as the commonalities (such as the historically large level of public capital in the post–World War II period, and the large decline in recent decades in high-income countries as well as in Russia or China, with important consequences for the distribution of private wealth and the rise of new forms of oligarchs).

Given the specific and context-heavy discussion of these multidimensional factors, does it still make sense to speak of "capital" as a single category? The fact that it is technically possible to add up all the market values of the different existing assets (to the extent that such market values are well defined, which is not always entirely clear) in order to compute the aggregate value of the capital stock K does not change anything about the basic multidimensional reality of assets and corresponding property relations. I attempt to show that this abstract operation can be useful for some purposes. In particular, by computing the ratio $\beta = K/Y$ between the aggregate market value of capital K and national income Y, one can compare the overall importance of capital wealth, private property, and public property in societies that are otherwise impossible to compare. For instance, one finds that in spite of all metamorphosis in the nature of assets and institutional arrangements, aggregate capital values—expressed relative to total national income—are in a number of countries approaching the levels observed in the patrimonial societies that flourished in the 18th–19th centuries and until World War I. I believe that this finding is interesting in itself. But it certainly does not alter the fact that a proper comparison of these different societies requires a careful separate analysis of the various asset categories and corresponding social and economic relations.

Inequality of Labor Income and Inequality of Wealth

Another way in which my analysis of capital and inequality is multidimensional is that throughout the book, I continually distinguish between the inequality of labor income and the inequality of capital ownership. Of course these two dimensions of inequality do interact in important ways: for example, rising inequality in labor earnings during a certain period of time might tend to fuel rising wealth concentration in following decades or generations. But the forces that drive income inequality and wealth inequality are largely different.

For instance, I point out in my book (particularly Chapters 8–9) that the rise of top income shares in the United States over the 1980–2010 period is due for the most part to rising inequality of labor earnings, which can itself be explained by a mixture of two groups of factors: 1) rising inequality in access to skills and to higher education over this time period in the United States, an evolution which might have been exacerbated by rising tuition fees and insufficient public investment; and 2) exploding top managerial compensation, itself probably stimulated by changing incentives and norms, and by large cuts in top tax rates (see also Chapter 14; Piketty, Saez, and Stantcheva 2014). More broadly, I argue (p. 243) that the mechanisms behind unequal incomes from labor "include the supply of and demand for different skills, the state of the educational system, and the various rules and institutions that affect the operation of the labor market and the determination of wages." This rise in labor earnings inequality in recent decades evidently has little to do with the gap $r - g$; indeed, it seems fairly difficult to find a logical way that $r - g$ could affect the inequality of labor income. Conversely, "[i]n the case of unequal incomes from capital, the most important processes involve savings and investment behavior, laws governing gift-giving and inheritance, the operation of real estate and financial markets, and so on" (p. 243).

In addition, the notions of top deciles or percentiles are not the same for the distributions of labor income and capital ownership. The use of deciles and percentages should be viewed as a language allowing for comparisons between societies that are otherwise impossible to compare, such as France in 1789 and China or the United States in 2014, in the same way as the aggregate capital-income ratio can be used to make comparisons. But in certain societies, the top shares of income and wealth might be highly correlated, while in other societies they may represent entirely different social hierarchies (as in traditional patrimonial societies). The extent to which these two dimensions of inequality differ gives rise to different representations and beliefs systems about social inequality, which in turn shape institutions and public policies affecting inequality dynamics.

The Dynamics of Wealth Inequality and the Role of $r > g$

Let me now try to clarify the role played by $r > g$ in my analysis of inequality dynamics. The rate of return on capital is given by r, while g measures the rate of economic growth. The gap between r and g is certainly not the only relevant mechanism for analyzing the dynamics of wealth inequality. As I explained in the previous sections, a wide array of institutional factors are central to understanding the evolution of wealth.

Moreover, the insight that the rate of return to capital r is permanently higher than the economy's growth rate g does not in itself imply anything about wealth inequality. Indeed the inequality $r > g$ holds true in the steady-state equilibrium of most standard economic models, including in representative-agent models where each individual owns an equal share of the capital stock.

For instance, consider the standard dynastic model where each individual behaves as an infinitely lived family and where the steady-state rate of return is well known to be given by the modified "golden rule" $r = \theta + \gamma g$ (where θ is the rate of time preference and γ is the curvature of the utility function). For example, if

$\theta = 3$ percent, $\gamma = 2$, and $g = 1$ percent, then $r = 5$ percent.[1] In this framework, the inequality $r > g$ always holds true, and this does not entail any implication about wealth inequality.

In a representative agent framework, what $r > g$ means is that in steady-state each family only needs to reinvest a fraction g/r of its capital income in order to ensure that its capital stock will grow at the same rate g as the size of the economy, and the family can then consume a fraction $1 - g/r$. For example, if $r = 5$ percent and $g = 1$ percent, then each family will reinvest 20 percent of its capital income and can consume 80 percent. Again, $r > g$, but this tells us nothing at all about inequality: this is simply saying that capital ownership allows the economy to reach higher consumption levels—which is really the very least one can ask from capital ownership.[2]

So what is the relationship between $r - g$ and wealth inequality? To answer this question, one needs to introduce extra ingredients into the basic model so that inequality arises in the first place.[3] In the real world, many shocks to the wealth trajectories of families can contribute to making the wealth distribution highly unequal (indeed, in every country and time period for which we have data, wealth distribution *within each age group* is substantially more unequal than income distribution, which is difficult to explain with standard life-cycle models of wealth accumulation; for a concise summary of the historical evidence on the extent of income and wealth inequality, see Piketty and Saez 2014). There are demographic shocks: some families have many children and have to split inheritances in many pieces, some have few; some parents die late, some die soon; and so on. There are also shocks to rates of return: some families make very good investments, others go bankrupt. There are shocks to labor market outcomes: some earn high wages, others do not. There are differences in taste parameters that affect the level of saving: some families consume

[1] Intuitively, in a model where everyone maximizes an infinite-horizon utility function $U = \int_{0 \leq t \leq +\infty} e^{-\theta t}\, u(c_t)$ (with $u(c) = c^{1-\gamma}/(1-\gamma)$), then $r = \theta + \gamma g$ is the unique rate of return to capital possible in the long-run for the following reason: it is the sole rate such that the agents are willing to raise their consumption at rate g, that is at the growth rate of the economy. If the return is higher, the agents prefer to postpone their consumption and accumulate more capital, which will decrease the rate of return; and if it is lower, they want to anticipate their consumption and borrow more, which will increase the rate of return.

[2] The inequality $r < g$ would correspond to a situation which economists often refer to as "dynamic inefficiency": in effect, one would need to invest more than the return to capital in order to ensure that one's capital stock keeps rising as fast as the size of the economy. In infinite horizon models with perfect capital markets, this cannot happen. In effect, $r < g$ would violate the transversality condition: the net present value of future resources would be infinite, so that rational agents would borrow infinite amounts in order to consume right away. However, in models with other saving motives, such as finite-horizon overlapping generation models, it is possible for $r < g$.

[3] In the dynastic model with no shock, there is no force generating inequality out of equality (or equality out of inequality), so that any initial level of wealth inequality (including full equality) can be self-sustaining, as long as the modified "golden rule" is satisfied. In effect, steady-state wealth inequality is exogenous and indeterminate, and does not depend on the gap $r - g$. Note however that the magnitude of the gap $r - g$ has an effect on the steady-state inequality of consumption and welfare in this basic model: for example, if $r - g$ is small, then high-wealth dynasties need to reinvest a large fraction of their capital income, so that they do not consume much more than low-wealth dynasties.

a lot more than a fraction $1 - g/r$ of their capital income and might even consume away the capital value and die with negligible wealth; others might reinvest a lot more than a fraction g/r and have a strong taste for leaving bequests and perpetuating large fortunes.

A central property of this large class of models is that for a given structure of shocks, the long-run magnitude of wealth inequality will tend to be magnified if the gap $r - g$ is higher. In other words, wealth inequality will converge towards a finite level in these models. The shocks will ensure that there is always some degree of downward and upward wealth mobility such that wealth inequality remains bounded in the long run. But this finite inequality level will be a steeply rising function of the gap $r - g$. Intuitively, a higher gap between r and g works as an amplifier mechanism for wealth inequality for a given variance of other shocks. To put it differently: a higher gap between r and g allows an economy to sustain a level of wealth inequality that is higher and more persistent over time (that is, a higher gap $r - g$ leads both to higher inequality and lower mobility).

More precisely, one can show that if shocks take a multiplicative form, then in the long run, the inequality of wealth will converge toward a distribution that has a Pareto shape for top wealth holders (which is approximately the form that we observe in real-world distributions and corresponds to relatively fat upper tails and a large concentration of wealth at the very top), and that the inverted Pareto coefficient (an indicator of top-end inequality) is a steeply rising function of the gap $r - g$.[4] This well-known theoretical result was established by a number of authors using various structures of demographic and economic shocks (see in particular Champernowne 1953; Stiglitz 1969). The logic behind this result and this "inequality amplification" impact of $r - g$ is presented in Chapter 10 of my book: for detailed references to this literature on wealth inequality, $r - g$, and Pareto coefficients see the online appendix to Chapter 10 of my book (available at http://piketty.pse.ens.fr/capital21c) and Piketty and Zucman (2015, section 5.4). These connections between $r - g$ and Pareto coefficients of steady-state wealth distributions are also explained very clearly in the review by Charles Jones in the present symposium.

In this class of models, relatively small changes in $r - g$ can generate very large changes in steady-state wealth inequality. For example, simple simulations of the model with binomial taste shocks show that going from $r - g = 2$ percent to $r - g = 3$ percent is sufficient to move the inverted Pareto coefficient from $b = 2.28$ to $b = 3.25$. This corresponds to a shift from an economy with moderate wealth inequality—say, with a top 1 percent wealth share around 20–30 percent, such as present-day Europe or the United States—to an economy with very high wealth

[4] A Pareto distribution means that above a certain wealth level z_0, the population fraction with wealth above z is given by $p(z) = p_0(z_0/z)^a$ (where a is a constant). A characteristic property of the Pareto distribution is that the ratio $b = \mathrm{E}(z|z > z')/z'$ between average wealth above some threshold z' and the level of the threshold z' is independent of z' and is equal to the inverted Pareto coefficient $b = a/(a - 1)$.

inequality, with a top 1 percent wealth share around 50–60 percent, such as Europe in the 18th–19th centuries and up until World War I.[5]

To summarize: the effect of $r - g$ on inequality follows from its dynamic cumulative effects in wealth accumulation models with random shocks, and the quantitative magnitude of this impact seems to be sufficiently large to account for very important variations in wealth inequality.

To reiterate, this argument does not imply that the $r - g$ effect is the only important force that matters in accounting for historical variations in wealth inequality. The variance of other shocks (particularly to rates of returns, which vary enormously across assets and individuals), as well the income and wealth profiles of saving rates, obviously matter a great deal. Most importantly, it is really the *interaction* between the $r - g$ effect and the institutional and public policy responses—including progressive taxation of income, wealth, and inheritance; inflation; nationalizations, physical destruction, and expropriations; estate division rules; and so on—which in my view, determines the dynamics and the magnitude of wealth inequality. In particular, if one introduces taxation into the basic model, then it follows immediately that what determines long-run wealth inequality and the steady-state Pareto coefficient is the gap $(1 - t)r - g$ between the net-of-tax rate of return and the growth rate.

In their contribution to this symposium, Acemoglu and Robinson present cross-country regression results between income inequality and $r - g$ and argue that $r - g$ does not seem to have much impact on inequality. However, I do not find these regressions very convincing, for two main reasons. First, income inequality is primarily determined by the inequality of labor income (which typically represents between two-thirds and three-quarters of total income), which as I noted above has nothing to do with $r - g$. It would make more sense to run such a regression with wealth inequality, but long-run wealth inequality series are available for a much more limited number of countries than income inequality series. In Chapter 12 of my book, I present wealth inequality series for only four countries (France, Britain, Sweden, and the United States), and the data are far from perfect. We do plan in the future to extend the World Top Incomes Database (WTID) into a World Wealth and Income Database (W2ID) and to provide homogenous wealth inequality series for all countries covered in the WTID (over 30 countries). But at this stage, we have to do with what we have.

[5] In the special case with saving taste shocks, the transition equation for normalized wealth $z_{ti} = w_{ti}/w_t$ (where w_{ti} is the wealth level of dynasty i at period t, and w_t is average wealth at period t) is given by: $z_{t+1i} = (s_{ti}/s) \cdot [(1 - \omega) + \omega \cdot z_{ti}]$, with $\omega = s \cdot e^{(r-g)H}$ (where s is the average saving taste parameter, s_{ti} is the taste parameter of dynasty t at period t, r and g are the annual rate of return and growth rate, and H is generation length). With binomial shocks with probability p, one can show that the inverted Pareto coefficient is given by $b = \log(1/p)/\log(1/\omega)$. See Piketty and Zucman (2015, section 5.4) for calibrations of this formula. In Atkinson, Piketty, and Saez (2011, figures 12–15, p. 50–55), we provide evidence on the long-run evolution of inverted Pareto coefficients for income distributions. See also the discussion in the online appendix to Chapter 10 of my book (available at http://piketty.pse.ens.fr/capital21c).

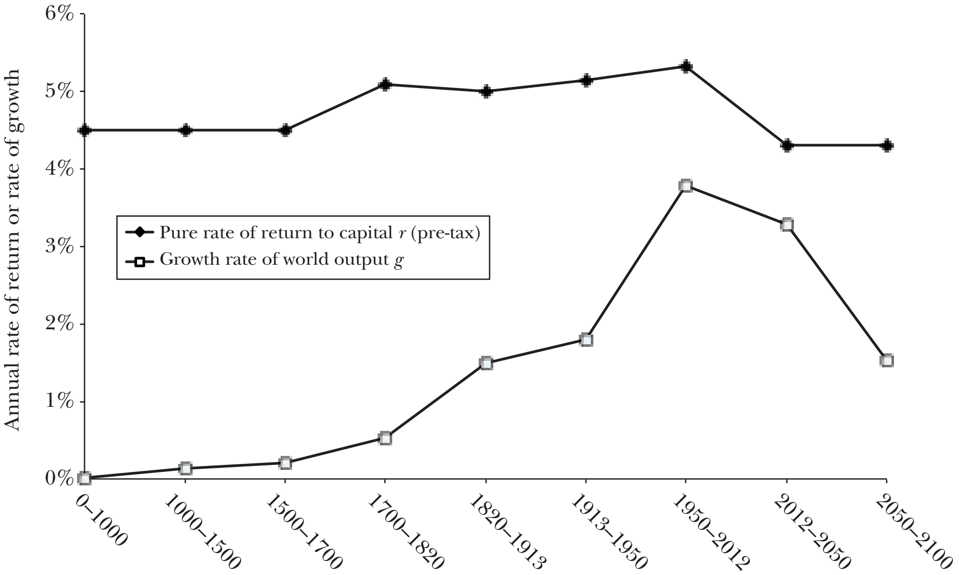
Second, the process of intergenerational accumulation and distribution of wealth is very long-run process, so looking at cross-sectional regressions between inequality and $r - g$ may not be very meaningful. One would need to introduce time lags, possibly over very long time periods: for example, one might use the average $r - g$ observed over 30 or 50 years. As I argue below, the broad correlations between $r - g$ and wealth inequality certainly seem to run in the right direction, both from a long-run (18th–19th versus 20th centuries) and international (Europe versus US) perspective. However, given the data limitations and the time-lag specification problems, I am not sure there is a lot to learn from running explicit cross-country regressions.

In my view, a more promising approach—on this issue as well as on many other issues—is a mixture of careful case studies and structural calibrations of theoretical models. Although we do not have many historical series on wealth inequality, they show a consistent pattern. Namely, we observe extremely high concentration of wealth in pretty much every European society in the 18th and 19th centuries up until World War I. In particular, in France, Britain, and Sweden, the top 10 percent wealth share was about 90 percent of total wealth (including a top 1 percent wealth share of around 60–70 percent) in the 19th century and at the very beginning of the 20th century. If anything, wealth inequality seems to have been rising somewhat during the 19th century and up until World War I—or maybe to have stabilized at very high levels around 1890–1910. Thus, in spite of the large changes in the nature of wealth during the 19th century—agricultural land as a form of wealth is largely replaced by real estate, business assets, and foreign investment—wealth inequality was as extreme in the modern industrial society of 1914 as it had been under France's *ancien regime* in 1789.

The most convincing explanation for the very high wealth concentration in these pre–World War I European societies seems to be the very large $r - g$ gap—that is, the gap between rates of return and growth rates during the 18th and 19th centuries. There was very little taxation or inflation up until 1914, so the gap $(1 - t)r - g$ was particularly high in pre–World War I societies, which in dynamic models of wealth accumulation with random shocks leads to very large wealth concentration. In contrast, following the large capital shocks of the 1914–1945 period—a time of physical destruction, periods of high inflation and taxation, and nationalizations—the after-tax, after-capital-losses rate of return precipitously fell below the growth rate after World War I. Figure 1 compares the pre-tax pure rate of return with growth rate g, while Figure 2 shows a post-tax, post-losses rate of return, including projections into the future.

This interpretation of the evidence is further confirmed by the detailed individual-level data collected in French inheritance archives since the time of the French Revolution (Piketty, Postel-Vinay, and Rosenthal 2006, 2014). We find that the more and more steeply increasing age-wealth profiles at high wealth levels in the 19th century and early 20th century can be well accounted for by a capitalization effect and a high gap between $(1 - t)r$ and g. This age–wealth pattern suddenly breaks down following the 1914–1945 capital shocks. The fact that US

Figure 1
Rate of Return versus Growth Rate at the World Level, from Antiquity until 2100



Source: Author (figure 10.9 from Piketty 2014). For more on sources and series, see http://piketty.pse.ens.fr/capital21c.
Note: The rate of return to capital (pre-tax) has always been higher than the world growth rate, but the gap was reduced during the 20th century, and might widen again in the 21st century.

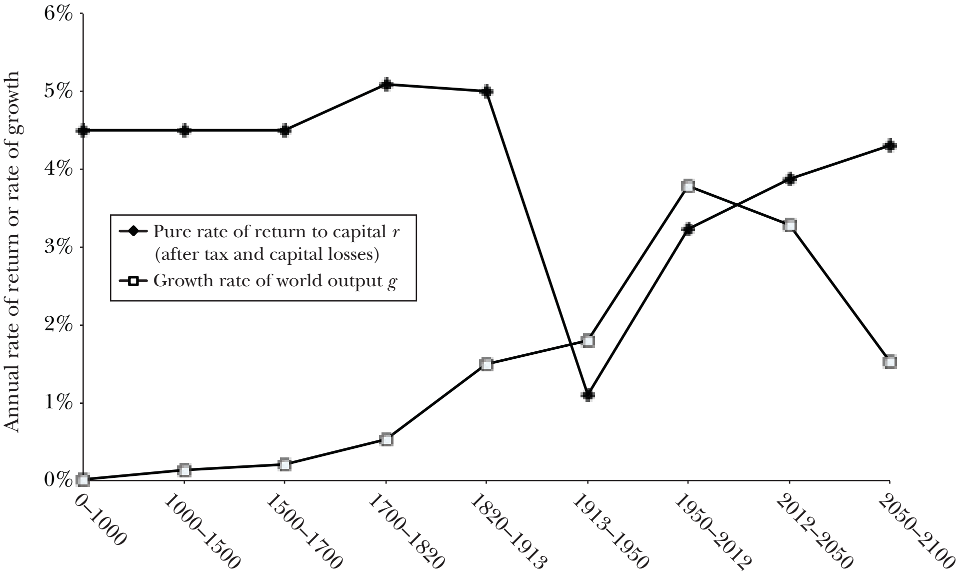
wealth concentration was significantly less than in Europe during the 19th century and up until World War I is also consistent with this model: growth rates were higher in the US economy, in particular due to higher population growth, thereby limiting the dynamic cumulative effects of the inequality amplification channel. Also, there had been less time for dynastic wealth concentration to arise in the US economy by the 19th century. This evidence is further reviewed in Chapters 10–11 of my book.

Data collection in French archives and in other countries will continue, and new data will certainly allow for better empirical tests of wealth accumulation models in the future. But at this stage, the best evidence we have suggests that $r > g$ is an important part of the explanation for the very high and persistent level of wealth concentration that we observe in most societies in the 18th–19th centuries and up until World War I.

What Will Be the Evolution of $r - g$ in the 21st Century?

A number of forces might lead to greater inequality of wealth in the 21st century, including a rise in the variance of shocks to demographic factors, rates of return,

Figure 2
After-Tax Rate of Return versus Growth Rate at the World Level, from Antiquity until 2100



Source: Author (figure 10.10 from Piketty 2014). For more on sources and series, see http://piketty.pse.ens.fr/capital21c.
Note: The rate of return to capital (after tax and capital losses) fell below the growth rate during the 20th century, and may again surpass it in the 21st century.

labor earnings, tastes for saving and bequests, and so on. Conversely, a reduction of the variance of these shocks could lead to a decline in wealth inequality. The gap between $(1 - t)r$ and g is certainly not the only determinant of steady-state wealth inequality. It is one important determinant, however, and there are reasons which might push toward a persistently high gap between the net-of-tax rate of return $(1 - t)r$ and the growth rate g in the 21st century—which might in turn lead to higher steady-state wealth inequality (other things equal). In my book, I particularly emphasize the following three potential forces: global tax competition to attract capital; growth slowdown and technical change; and unequal access to high financial returns (Chapters 10–12). Here, I restate and sharpen some of the main arguments.

As international competition intensifies to attract investment, it is plausible that capital taxes will fall, as they have already been doing in many countries in the last few decades. By capital taxes, I include both corporate profit taxation and wealth and inheritance taxes. But of course, the ultimate effect of tax competition will depend on the institutional response. If a sufficiently large number of countries manage to better coordinate to establish a common corporate tax on large corporations and a reliable system of automatic transmission of information of cross-border financial assets, then the effective capital tax rate might rise, in which case $(1 - t)r$ will decline, and so will steady-state wealth inequality. Ultimately, the outcome depends on the institutional response. Indeed, recent research indicates that better international fiscal coordination is difficult but by no means impossible (Zucman 2014).

Note also that a decline in capital tax rates and a rise in the after-tax rate of return $(1 - t)r$ might in principle induce an increase in saving rates and capital accumulation, thereby leading to a decline in the marginal product of capital which could partly undo the rise in the after-tax rate of return. Indeed, in the example mentioned earlier of the benchmark infinite-horizon dynastic model with no shock and a representative agent, in the long run, the after-tax rate of return to capital has to follow the rule $(1 - t)r = \theta + \gamma g$. In this case, the tax cut leads to a savings response that ultimately moves the rate of return completely back to its earlier level. However, this outcome only arises due to an extreme and unrealistic assumption: namely, the long-run elasticity of saving and capital accumulation with respect to after-tax rate of return is infinite in such a model. In more realistic dynamic models of capital accumulation where this elasticity is positive but not infinite, a decline in capital tax will lead to a net increase in the after-tax rate of return in the long run.[6]

The effect of a growth slowdown on $r - g$ and on the long-run dynamics of wealth inequality is more complicated to analyze. In the historical data, the pre-tax rate of return r seems to display little historical variation, so that $r - g$ definitely appears to be smaller than when the growth rate is higher, as illustrated earlier in Figure 1. This would tend to support the view that lower growth rates in the 21st century (in particular due to the projected decline of population growth) are likely to contribute to a rise of $r - g$.[7]

From a theoretical perspective, however, the effect of a decline in the growth rate g on the gap $r - g$ is ambiguous: it could go either way, depending on how a change in g affects the long-run rate of return r. This depends on a mixture of forces, including saving behavior, multisector technological substitution, bargaining power, and institutions. Let me summarize the main arguments (see Chapters 5–6 of my book for a more thorough analysis; see also the discussion of this point by Jones in this symposium). Generally speaking, a lower g, due either to a slowdown of population and/or productivity growth, tends to lead to a higher steady-state capital–output ratio $\beta = K/Y$, and therefore to lower rates of return to

[6] For a class of dynamic capital accumulation models with finite long-run elasticities of saving with respect to after-tax rates of returns, and for a study of corresponding socially optimal tax rates on capital, see Piketty and Saez (2013). One of the important findings is that the optimal tax rate is an increasing function of $r - g$ (due in particular to the inequality effect of $r - g$).

[7] This conclusion largely depends on the way the corrected rates of return reported on Figure 1 were constructed: specifically, the rates of return implied by conventionally measured capital shares are generally very large in high-growth, reconstruction periods. Chapter 6 of my book offers a discussion as to why such high returns might include substantial entrepreneurial labor input and should therefore be corrected downwards; such corrections are highly uncertain, however.

capital r (for given technology). The key question is whether the fall in r is smaller or larger than the fall in g. There are, in my view, good reasons to believe that r might fall less than the fall in g, but this issue is a complex one.

In the benchmark dynastic model, the steady-state β rises as g declines, and the rate of return $r = \theta + \gamma\, g$ drops. Whether $r - g$ rises or declines as g declines depends entirely on whether the curvature of the utility function γ is smaller or larger than one. However this model does not seem to be particularly realistic empirically, so this may not be the best way to look at the problem. Note that the dynastic model can be viewed as a special case of the general Harrod–Domar–Solow steady-state formula $\beta = s/g$. In effect, in the steady-state of the dynastic model, the (net-of-depreciation) saving rate $s = s(g)$ rises moderately with g, so that $\beta = s(g)/g$ is a declining function of g.[8]

If one instead assumes a fixed, exogenous saving rate s, then the steady-state capital output ratio $\beta = s/g$ will rise even more strongly as g declines. With perfect competition and a constant-elasticity-of-scale production function, whether the resulting decline in r will more than compensate for a decline in g depends (among other things) on the value of the elasticity of substitution. With high substitutability between capital and labor (which might happen because of the rise of new capital-intensive technologies such as robots of various sorts), the rate of return will decline relatively little as β rises, so that $r - g$ will be higher with lower g.[9] In recent decades, the rise in the capital–income ratio β came together with a rise in the net-of-depreciation capital share α, which in a one-good model with perfect competition implies an elasticity of substitution higher than one. However, the one-good, perfect competition model is not a very satisfactory model, to say the least. In practice, the right model to think about rising capital–income ratios and capital shares is a multisector model (with a large role played by capital-intensive sectors such as real estate and energy, and substantial movements in *relative* prices) with important variations in bargaining power over time (see Chapters 5–6; see also Karababounis and Neiman 2014 about the role played by the declining relative price of equipment). In particular, intersectoral elasticities of substitution combining supply and demand forces can arguably be much higher than within-sector capital–labor elasticities.

Note also there is, of course, no reason why the net-of-depreciation saving rates should be viewed as a constant. What I have in mind is an intermediate model (intermediate between the dynastic model and the exogenous saving model), with a relatively low elasticity of saving behavior with respect to r over a large range of middle returns (say, from 3 to 6 percent) and a much higher elasticity if rates of return take very low or very high values. In particular, if g becomes increasingly

[8] With a Cobb–Douglas production function $Y = \mathrm{F}(K, L) = K^{\alpha}L^{1-\alpha}$, the long-run capital–output ratio is given by $\beta = \alpha/r = \alpha/(\theta + \gamma g) = s(g)/g$, with $s(g) = \alpha g/r = \alpha g/(\theta + \gamma g)$. See Piketty and Zucman (2014).

[9] With $Y = \mathrm{F}(K, L) = [aK^{(\sigma-1)/\sigma} + (1 - a)L^{(\sigma-1)/\sigma}]^{\sigma/(\sigma-1)}$, the marginal productivity of capital is given by: $r = \mathrm{F}_K = a(Y/K)^{1/\sigma} = a\beta^{-1/\sigma}$.

close to zero, then it is clear that $\beta = s/g$ will not go to infinity: otherwise the rate of return would go to zero, and most agents would probably stop saving. In historical periods with very low growth rates (such as in pre-industrial societies), we observe large capital–income ratios, but not infinite β. As pointed out by Jones (in this symposium) and others, another obvious reason why β will not go to infinity is that depreciation would then become enormous. This intermediate model might explain why the rate of return seems to display limited systematic variations in the long run: it is roughly stable within a given range, which one might interpret as an interval of psychologically plausible time preference parameters.

Yet another way to explain why the rate of return appears to be relatively stable in the long run is the following. Pure economic reasoning tends to imply that higher growth leads to higher returns. But high growth periods arguably require more entrepreneurial labor in order to reallocate capital continually and thus to benefit from higher returns (in other words, measured rates of return must be corrected downwards in order to take into account mismeasured labor input in high-growth societies). Conversely, measured rates of returns might be closer to pure returns in low-growth societies (where it is relatively easier to be a rentier, since capital reallocation requires less attention). This is the interpretation that I favor in the book; indeed, the historical estimates of rates of return in the book (those given above and in Chapter 6 of the book) are largely built upon this assumption.

If we combine all these different effects, it is clear however that there is no general, universal reason why $r - g$ should increase as g declines: it could potentially go either way. Historical evidence and new technological developments suggest that it should increase (and I tend to favor this conclusion), but I fully agree that this remains relatively uncertain.

Finally, the last reason (and arguably the most important one) why $r - g$ might be high in the 21st century is due to unequal access to high financial returns. That is, even though the gap between the average rate of return r and the growth rate g is not particularly high, it could be that large potential financial portfolios have access to substantially higher returns than smaller ones. In the book, I present evidence suggesting that financial deregulation might have contributed to such an evolution (Chapter 12). For example, according to *Forbes* rankings, the wealth of top global billionaires seem to be rising much faster than average wealth, as shown in Table 1. This evolution cannot continue for too long, unless one is ready to accept an enormous increase in the share of world wealth belonging to billionaires (and a corresponding decline in the share going to the middle class). Also, larger university endowments tend to obtain substantially higher returns, as shown in Table 2 (and the data presented by Saez and Zucman 2014 on nonprofit foundations indicates a similar pattern). This data is clearly imperfect and too incomplete to prove the general theme of unequal access to high returns. But given that even small changes in $r - \mathrm{g}$ can have large amplifying effects on changes in wealth inequality, this effect is potentially important.

Overall, there remains substantial uncertainty about how far wealth inequality might rise in the 21st century, and we need more transparency and better

Table 1
The Growth Rate of Top Global Wealth, 1987–2013

	Average real growth rate per year (after deduction of inflation) 1987–2013
For top 1/(100 million) highest wealth-holders (about 30 adults out of 3 billion in 1980s, and 45 adults out of 4.5 billion in 2010s)	6.8%
For top 1/(20 million) highest wealth-holders (about 150 adults out of 3 billion in 1980s, and 225 adults out of 4.5 billion in 2010s)	6.4%
For average world wealth per adult	2.1%
For average world income per adult	1.4%
For world adult population	1.9%
For world GDP	3.3%

Source: Table 12.1 from Piketty (2014). For more information, see http://piketty.pse.ens.fr/capital21c.
Notes: Between 1987 and 2013, the highest global wealth fractiles have grown at 6–7 percent per year, versus 2.1 percent for average world wealth and 1.4 percent for average world income. All growth rates are net of inflation (2.3 percent per year between 1987 and 2013).

Table 2
The Return on the Capital Endowments of US Universities, 1980–2010

	Average real annual rate of return (after deduction of inflation and all administrative costs and financial fees) 1980–2010
For all universities (850)	8.2%
Harvard-Yale-Princeton	10.2%
Endowments higher than 1 billion $ (60)	8.8%
Endowments between 500 million and 1 billion $ (66)	7.8%
Endowments between 100 and 500 million $ (226)	7.1%
Endowments less than 100 million $ (498)	6.2%

Source: Table 12.2 from Piketty (2014). For more information, see http://piketty.pse.ens.fr/capital21c.
Notes: Between 1980 and 2010, US universities earned an average real return of 8.2 percent on their capital endowments, and even more for higher endowments. All returns reported here are net of inflation (2.4 percent per year between 1980 and 2010) and of all administrative costs and financial fees.

information about wealth dynamics. In my view, one main benefit of a progressive wealth tax is that it would produce better information regarding the size and evolution of different wealth groups such that the wealth tax could be adapted in the future on the basis of this better information. I agree with the argument by Kopczuk in this symposium that the data sources about the distribution of wealth that we have at our disposal are insufficient. At this stage, however, it seems to me that the method that infers wealth from the resulting income flows, the income capitalization method developed by Saez and Zucman (2014), produces probably the most reliable estimates we have, and these estimates show substantial recent rise in US wealth inequality—indeed, a higher rise than what I report in my book. In particular, Saez and Zucman find increasing concentration of capital income for all asset income categories (including dividend and interest, which cannot easily be contaminated by labor income considerations). Finally, the Saez and Zucman findings are consistent with the finding from the *Forbes* rankings that the wealth of top wealth-holders is rising much faster than average wealth. However, it is clear that these evolutions remain relatively uncertain. In my view, this makes the lack of transparency about wealth dynamics—largely due to the absence of a comprehensive wealth tax and the limitations of international coordination—particularly problematic.

Toward a New Historical and Political Economy Approach to Institutions

In my book *Capital in the Twenty-First Century*, I attempt to develop a new historical and political economy approach to the study of institutions and inequality dynamics. Economic forces such as the supply and demand for skills, wage bargaining models, or the effect of $r - g$ on wealth dynamics, also play a role. But ultimately, what really matters is the interaction between economic forces and institutional responses, particularly in the area of educational, labor, and fiscal institutions. Given my strong emphasis on how institutions and public policies shape the dynamics of income and wealth inequality, it is somewhat surprising that Acemoglu and Robinson argue in their contribution to this symposium that I neglect the role of institutions. It seems to me that we disagree less intensively than what they appear to believe, and that the well-known academic tendency to maximize product differentiation might be at work here.

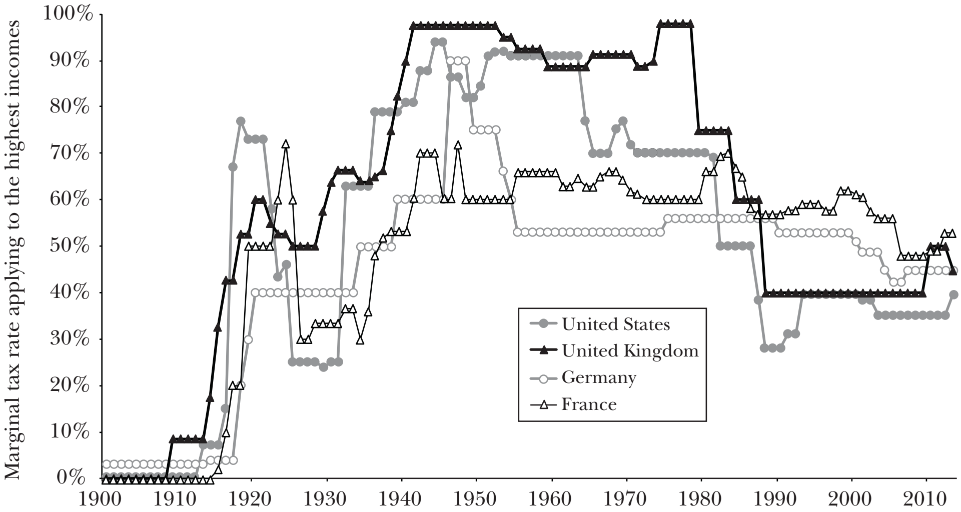
It is also possible that some of the confusion comes from the fact that we do not have exactly the same approach to the study of "institutions." However I believe that our approaches are broadly consistent and complementary to one another: they differ in terms of specific institutional content, as well as in time and geographical scope, more than in substance. In some of their earlier work, Acemoglu and Robinson mostly focused upon a relatively specific institution, namely the protection of property rights. In their fascinating book *Why Nations Fail*, they develop a broader view of institutions and stress the distinction between "inclusive" and "extractive" institutions. This broad concept might certainly include the type of institutions and policies on which I focus upon, including progressive taxation of income, wealth, and inheritance, or the modern welfare state. I must confess, however, that seeking to categorize institutions with broad terms like these strikes me as maybe a little too abstract, imprecise, and ahistorical.

I believe that institutions like the welfare state, free education, or progressive taxation, or the effects of World War I, the Bolshevik revolution, or World War II on inequality dynamics and institutional change, each need to be analyzed in a precise and concrete manner within the historical, social, and political context in which they develop. While Acemoglu and Robinson (2012) in their earlier book take a very long-run perspective on the history of the planet (from prehistoric times to the "great discoveries" and the formation of the modern world), I tend to focus on the historical periods and countries on which I was able to collect systematic data, that is, on the 18th, 19th, and especially the 20th centuries (an important period indeed for the formation of the modern social and fiscal state).

My approach to institutions emphasizes the role of political conflict in relation to inequality. In particular, wars and revolutions play a large role in my account of inequality dynamics and institutional change in the 20th century. Of course, steady democratic forces caused by the extension of suffrage also played an important role in the rise of more inclusive social, educational, and fiscal institutions during the 19th and 20th centuries. But many of the most important changes did not come simply from the steady forces of peaceful electoral democracy: rather, specific historical events and political shocks often played an important role. For example, there is little evidence of a natural movement toward more progressive taxation until the violent military, political, and ideological shocks induced by World War I (see Figure 3). Belief systems and collective representations about social inequality and the role of government were deeply affected by World War I and the rise of communism, as they were by the Great Depression, World War II, and then, at the end of the 20th century, by the stagflation of the 1970s and the fall of the Soviet Union.

It is particularly interesting to note that until 1914, the French elite often justified its strong opposition to the creation of a progressive income tax by referring to the principles of the French Revolution. In the view of these elites, France had become equal after 1789 thanks to the end of aristocratic privileges and the development of well-protected property rights for the entire population. Because everybody had been made equal in their ability to hold property, there was no need for progressive taxation (which would be suitable for aristocratic Britain, the story went, but not for republican France). What I find particularly striking in this pre-1914 debate is the combination of strong beliefs in property-rights-centered institutions and an equally strong denial of high inequality. In my book, I try to understand what we can learn from the fact that wealth inequality was as large in France in 1914 as in 1789, and also from the fact that much of the elite was trying to deny this. I believe there are important implications for the current rise in wealth and income inequality and the current attempts to minimize or deny that they are occurring. Then as now, when various shocks are tending to push wealth (and income) inequality higher at

Figure 3
Top Income Tax Rates, 1900–2013



Source: Author (figure 14.1 from Piketty 2014). For more on sources and series, see http://piketty.pse.ens.fr/capital21c.
Note: The top marginal tax rate of the income tax (applying to the highest incomes) in the United States dropped from 70 percent in 1980 to 28 percent in 1988.

a time when $r - g$ is at sustained high levels, the result can be a concentration of wealth that is high in historical terms.

Of course, I am not arguing that it will always take wars, revolutions, and other disruptive or violent political shocks to make institutional changes happen. In the case of early 20th century Europe, one can certainly argue that extreme inequality contributed to the high social tensions of the time and the rise of nationalism. But beliefs systems and resulting perceptions and policies can also be affected by peaceful public discussion. However we should not take this for granted. It is important to recognize the role of political conflict in the history of inequality and institutional change. It often took major fights to deliver change in the past, and it is not impossible that it will be the same in the future.

More generally, one of the lessons that I draw from this work is that the study of inequality dynamics and institutional change are intimately related. The development of stable institutions and the construction of a legitimate and centralized government are closely linked to the way different societies are able to address the issue of social inequality in a peaceful and orderly manner. In order to put institutions back at the center of economics, I believe that it is also necessary to put the study of distribution back at the center of economics. Institutions do not arise out of harmonious societies populated by representative agents; they arise out of unequal societies and out of conflict. This is again an issue on which the approaches developed by Acemoglu and Robinson and myself are broadly consistent and complementary.

Finally, let me conclude by making clear that my historical and political approach to inequality and institutions should be viewed as highly exploratory and incomplete. In particular, I suspect that new social movements and political mobilizations will give rise to institutional change in the future, but I do not pursue this analysis much further. As I look back at my discussion of future policy proposals in the book, I may have devoted too much attention to progressive capital taxation and too little attention to a number of institutional evolutions that could prove equally important, such as the development of alternative forms of property arrangements and participatory governance. One central reason why progressive capital taxation is important is that it can also bring increased transparency about company assets and accounts. In turn, increased financial transparency can help to develop new forms of governance; for instance, it can facilitate more worker involvement in company boards. But these other institutions also need to be analyzed on their own terms.

The last chapter of my book concludes: "Without real accounting and financial transparency and sharing of information, there can be no economic democracy. Conversely, without a real right to intervene in corporate decision-making (including seats for workers on the company's board of directors), transparency is of little use. Information must support democratic institutions; it is not an end in itself. If democracy is someday to regain control of capitalism, it must start by recognizing that the concrete institutions in which democracy and capitalism are embodied need to be reinvented again and again" (p. 570). I do not push this line of investigation much further, which is certainly one of the major shortcomings of my work. Together with the fact that we still have too little data on historical and current patterns of income and wealth, these are key reasons why my book is at best an introduction to the study of capital in the 21st century.

References

Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York: Crown Publishers.

Atkinson, Anthony B., and Alan J. Harrison. 1978. *Distribution of Personal Wealth in Britain.* Cambridge University Press.

Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez. 2011 "Top Incomes in the Long Run of History." *Journal of Economic Literature* 49(1): 3–71.

Champernowne, D. G. 1953. "A Model of Income Distribution." *Economic Journal* 63(250): 318–51.

Karabarbounis, Loukas, and Brent Neiman. 2014. "Capital Depreciation and Labor Shares around the World: Measurement and Implications." NBER Working Paper 20606.

Kuznets, Simon. 1953. *Shares of Upper Income Groups in Income and Savings.* National Bureau of Economic Research.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century.* Harvard University Press.

Piketty, Thomas, Gilles Postel-Vinay, and Jean-Laurent Rosenthal. 2006. "Wealth Concentration in a Developing Economy: Paris and France, 1807–1994." *American Economic Review* 96(1): 236–56.

Piketty, Thomas, Gilles Postel-Vinay and Jean-Laurent Rosenthal. 2014. "Inherited vs. Self-Made Wealth: Theory and Evidence from a Rentier Society (1872–1927)." *Explorations in Economic History* 51(1): 21–40.

Piketty, Thomas, and Emmanuel Saez. 2013. "A Theory of Optimal Inheritance Taxation." *Econometrica* 81(5): 1851–86.

Piketty, Thomas, and Emmanuel Saez. 2014. "Inequality in the Long Run." *Science* 344(6186): 838–43.

Piketty, Thomas, Emmanuel Saez, and Stefanie Stantcheva. 2014. "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities." *American Economic Journal: Economic Policy* 6(1): 230–71.

Piketty, Thomas, and Gabriel Zucman. 2014. "Capital is Back: Wealth–Income Ratios in Rich Countries, 1700–2010." *Quarterly Journal of Economics* 129(3): 1255–1310.

Piketty, Thomas, and Gabriel Zucman. 2015. "Wealth and Inheritance in the Long Run." In *Handbook of Income Distribution*, vol. 2, edited by A. Atkinson and F. Bourguignon, pp. 1303–68. Elsevier.

Saez, Emmanuel, and Gabriel Zucman. 2014. "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data." NBER Working Paper 20625.

Stiglitz, Joseph E. 1969. "Distribution of Income and Wealth Among Individuals." *Econometrica* 37(3): 382–97.

Zucman, Gabriel. 2014. "Taxing Across Borders: Tracking Personal Wealth and Corporate Profits." *Journal of Economic Perspectives* 28(4): 121–48.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 89–114

The Superiority of Economists[†]

Marion Fourcade, Etienne Ollion, and Yann Algan

There exists an implicit pecking order among the social sciences, and it seems to be dominated by economics. For starters, economists *see themselves* at or near the top of the disciplinary hierarchy. In a survey conducted in the early 2000s, Colander (2005) found that 77 percent of economics graduate students in elite programs agree with the statement that "economics is the most scientific of the social sciences." Some 15 years ago, Richard Freeman (1999, p. 141) speculated on the origins of such a conviction in the pages of this journal. His assessment was candid: "[S]ociologists and political scientists have less powerful analytical tools and know less than we do, or so we believe. By scores on the Graduate Record Examination and other criteria, our field attracts students stronger than theirs, and our courses are more mathematically demanding."

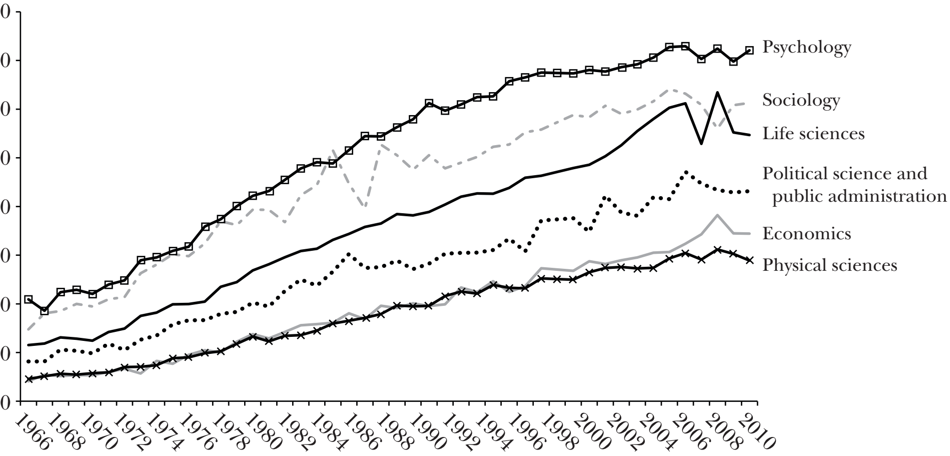
At first glance, the academic labor market seems to confirm the natives' judgment about the higher status of economists. They are the only social scientists to have a "Nobel" prize, thanks to a grant from the Bank of Sweden to the Nobel foundation. Economists command some of the highest levels of compensation in American arts and science faculties according to Bureau of Labor Statistics data. In fact, they "earn more and have better career prospects" than physicists and

■ *Marion Fourcade is Professor of Sociology, University of California, Berkeley, California, and Associate Fellow at the Max Planck-Sciences Po Center, Sciences Po, Paris, France. Etienne Ollion is Research Fellow at the Centre National de la Recherche Scientifique, University of Strasbourg, France. Yann Algan is Professor of Economics, Sciences Po, Paris, France. The authors' email addresses are fourcade@berkeley.edu, ollion@unistra.fr, and yann.algan@sciencespo.fr.*

[†] To access the Appendix and Data Appendix, visit
http://dx.doi.org/10.1257/jep.29.1.89

doi=10.1257/jep.29.1.89

Figure 1
Percentage of Doctorates Awarded to Women in Selected Disciplines, 1966–2011



Source: US National Center for Education Statistics, Integrated Post-secondary Education Data System Completion Survey.

mathematicians (as Freeman wrote); only computer scientists and engineers do better. Unlike many academics in the theoretical sciences and humanities, many prominent economists have the opportunity to obtain income from consulting fees, private investment and partnerships, and membership on corporate boards. For instance, Weyl (forthcoming) provides some suggestive evidence that 40 percent of the income of economic authors in the fields of finance and industrial organization comes from consulting activities, either with business (finance) or government (IO). In 2010, the movie documentary *Inside Job* exposed the lucrative and possibly complacent relations between some of the field's most distinguished members and the financial nebulae around Wall Street.

This much better financial position of economists, particularly in top universities, combined with the discipline's emphasis on mastering quantitative reasoning (widely interpreted as a sign of higher intellectual capabilities) certainly stands behind the often dismissive attitude of economists toward the other, less-formal social sciences. But there are other reasons for the distant relations among social scientists. First, the fields differ in their social composition. Self-selection into various disciplines is heavily patterned by social attributes. For instance, economics, like physics or philosophy but in sharp contrast to sociology, is a very male-dominated field (see Figure 1). Thus, cross-disciplinary relations are inevitably permeated by broader patterns of gender difference, stratification, and inequality. And while we do not have good comparative data on the social origins of social scientists in the United States (but see Bourdieu 1984 and Lebaron 2000 on France), we may posit that disparities in the present material conditions of the different fields generate important disparities in lifestyle and worldviews as well as relational strains between them.

Second, the social sciences have experienced fast demographic growth since World War II, which has produced internal differentiation and hyper-specialization. (Abbott 2001; Frank and Gabler 2006). This process has obscured the common heritage—the fact that history and moral philosophy gave birth to political economy back in the nineteenth century (see Haskell 1977; Ross 1991, for a US-focused discussion), while American sociology arose partly from within economics in the early part of the twentieth century (Young 2009).

In this essay, we explore the shifting relationship between economics and the other social sciences in four specific dimensions. First, we document the relative insularity of economics and its dominant position within the network of the social sciences in the United States. Though all disciplines are in some way insular—a classic consequence of the heightening of the division of academic labor (Jacobs 2013)—this trait peculiarly characterizes economics. Second, we document the pronounced hierarchy that exists within the discipline, especially in comparison with other social sciences. The authority exerted by the field's most powerful players, which fosters both intellectual cohesiveness and the active management of the discipline's internal affairs, has few equivalents elsewhere. Third, we look at the changing network of affiliations of economics over the post-World War II period, showing in particular how transformations within higher education (most prominently the rise of business schools) and the economy have contributed to a reorientation of economics toward business subjects and especially finance. Finally, we provide a few insights into the material situation, worldviews, and social influence of economists, which also set them apart from their academic peers. Taken together, these traits help to define and account for the intellectual self-confidence of economists and in turn for their assertive claims on matters of public policy.

When we refer to the "superiority of economists," our *double entendre* has both a descriptive and an explanatory purpose. Economics occupies a unique position among academic disciplines. It is characterized by far-reaching scientific claims linked to the use of formal methods; the tight management of the discipline from the top down; high market demand for services, particularly from powerful and wealthy parties; and high compensation. This position of social superiority also breeds self-confidence, allowing the discipline to retain its relative epistemological insularity over time and fueling a natural inclination towards a sense of entitlement. While the imperialistic expansion of economics into aspects of social science that were traditionally outside the economic canon has spurred some engagement with noneconomics scholarship, the pattern of exchange remains deeply asymmetrical, causing resentment and hostility in return. And while economists' unique position gives them unusual power to accomplish changes in the world, it also exposes them more to conflicts of interests, critique, and mockery when things go wrong.

Insularity

The intellectual trajectories of the social science disciplines have diverged importantly during the post-World War II period. Economics has changed since its continental youth. It left history behind and turned to the paradigmatic fields in the natural sciences, such as physics, for emulation (Mirowski 1989). Unlike their more literary forerunners, modern-day economists attribute their intellectual standing and autonomy to their reliance on precisely specified and parsimonious models and measures. They see the field's high technical costs of entry and its members' endeavors to capture complex social processes through equations or clear-cut causality as evidence of the discipline's superior scientific commitments, vindicating the distance from and the lack of engagement with the more discursive social sciences. In a prominent example, Lazear (2000, pp. 99–100) writes: "The ascension of economics results from the fact that our discipline has a rigorous language that allows complicated concepts to be written in relatively simple, abstract terms. The language permits economists to strip away complexity. Complexity may add to the richness of description, but it also prevents the analyst from seeing what is essential." An eminent professor echoed this view when he described, this time critically, the narrow epistemological demands of his discipline (interviewed by Fourcade 2009, p. 91): "You are only supposed to follow certain rules. If you don't follow certain rules, you are not an economist. So that means you should derive the way people behave from strict maximization theory. . . . The opposite [to being axiomatic] would be arguing by example. You're not allowed to do that. . . . There is a word for it. People say 'that's anecdotal.' That's the end of you if people have said you're anecdotal . . . [T]he modern thing [people say] is: 'it's not identified.' God, when your causality is not identified, that's the end of you."

For much of the post-World War II period, flexing one's mathematical and statistical muscles and stripping down one's argument to a formal and parsimonious set of equations was indeed the main path to establishing scientific purity in economics. With the empirical revolution in the 1990s and 2000s, this function has shifted toward a hard-nosed approach to causality focused on research design and inference and often extolling the virtues of randomly controlled trials (for example, Angrist and Pischke 2010). Although this move has not escaped criticism (for example, see Leamer 2010 and Sims 2010 in this journal), it represents a significant departure from the now disparaged over-theoretical orientations of the 1970s and 1980s. The shift toward applied microeconomics, while very real, has not dramatically broadened the network of interdisciplinary connections however. To be sure, economists have started to consider topics that are more traditionally associated with sociology, political science, and psychology—from political institutions to family structure, neighborhood effects, peer effects, or (as of late) social mobility. Yet cross-disciplinary citation patterns continue to offer evidence of the field's relative insularity. Of course, one of the most remarkable facts about US social science (continental Europe tends to be more ecumenical) is the extent to which *all* its constituent disciplines work in relative isolation from each other: economics, sociology, political science, and psychology all have high percentages of intradisciplinary citations. But even so, economics stands out markedly, with 81 percent of within-field citations in 1997—against 52 percent for sociology, 53 percent for anthropology, and 59 percent for political science (Jacobs 2013, p. 82, who uses the *NSF Science and Engineering Indicators 2000*, online appendix 6–54, based on a sample of the most cited journals in each field).

There are several reasons for the insularity of economics, most importantly the different epistemological cultures of the various social science disciplines and the power inequalities between them. First, the theory of action that comes with economists' analytical style is hardly compatible with the basic premise of much of the human sciences, namely that social processes shape individual preferences (rather than the other way around). In economics, by contrast, "de gustibus non est disputandum" (Stigler and Becker 1977)[1]: preferences are "usually assumed to be fixed" (Baron and Hannan 1994, p. 1116).[2] Second, the qualitative methods that underpin the work of many interpretive social scientists often do not square well with economists' formal aspirations, with their views on causality, or with their predilection for methodological and theoretical precision over real-world accuracy. Third, even when the substantive terrains overlap, the explicit or implicit pecking order between the disciplines often stands in the way of a desirable form of intellectual engagement.

Examining the structure of interdisciplinary citations in detail reveals sharp differences across disciplines. Surveying academic journals from 1995 to 1997, Pieters and Baumgartner (2002) found sharply asymmetric flows between economics and the other social sciences. Our analysis of citations in flagship journals for economics, sociology, and political science over the period from 2000 to 2009 confirms this pattern. As shown in Table 1, articles in the *American Political Science Review* cite the top 25 economics journals more than five times as often as the articles in the *American Economic Review* cite the top 25 political science journals. The asymmetry is even starker with regard to the *American Sociological Review.* While only 2.3 percent of the sociologists' citations go to their economic colleagues (often in a critical fashion, arguably), just 0.3 percent of economists' citations go to sociologists (again only taking into account the top 25 journals in each discipline). Citation data are, of course, likely to be biased downwards because sociology and political science tend to cast their citation networks more broadly overall and because of the role of books (which we do not account for) in those fields. Even so, it is worth pondering these asymmetrical patterns, especially since the discrepancy is so large and other sources of evidence all point in the same direction. A targeted comparison of citations to important figures in sociology and economics who deliberately engaged the other discipline shows this well. French sociologist Pierre Bourdieu, the top-cited name in US sociology today, received a single mention in the *AER* during the 2000s

[1] "In matters of taste, there can be no disputes."

[2] In the last 10–15 years however, a few economists have taken a more active interest in the formation of preferences. For examples, see Bowles (1998) and Fehr and Hoff (2011).

Table 1

Citations from the Flagship Journal to Articles Published in the 25 Top Journals in Each Discipline, 2000–2009

(as a percentage of total citations in each journal)

	Cited journals (% of all references)			
Citing journal	*Top 25 economics journals*	*Top 25 political science journals*	*Top 25 sociology journals*	*Total number of papers/citations from this journal*
American Economic Review	40.3%	0.8%	0.3%	907/ 29,958
American Political Science Review	4.1%	17.5%	1.0%	353/ 19,936
American Sociological Review	2.3%	2.0%	22%	399/ 23,993

Source: Compiled by the authors from the electronic Institute for Scientific Information's *Web of Social Science*. The high number of papers and cites in the *AER* is due to the Papers and Proceedings. We also looked at this data without the Papers and Proceedings. The patterns are not significantly different.

(against 60 times in the *ASR*), while Gary Becker was reaping 41 citations in the *ASR* (106 in *AER*). During the same period Max Weber and Mark Granovetter received four mentions each in the *AER*, but James Heckman was cited 25 times in the *ASR* by sociologists, and Oliver Williamson, 13.[3]

From the vantage point of sociologists, geographers, historians, political scientists, or even psychologists, economists often resemble colonists settling on their land—an image reinforced by some economists' proud claims of "economic imperialism" (Lazear 2000). Lured by the prospect of a productive crop, economists are swift to probe the new grounds. They may ask for guidance upon arrival, even partner-up with the locals (with whom they now often share the same data). But they are unlikely to learn much from them, as they often prefer to deploy their own techniques.[4] And in some cases, the purpose has been simply to set the other disciplines straight (Nik-Khah and van Horn 2012). Under the influence, notably, of Chicago price theory, the dominant economic paradigm has successfully conquered a segment of political science, law, accounting, and (for a while) sociology under the label of rational choice theory —thus explaining, in part, the directionality of the citation patterns observed above.

Opinion surveys further confirm this analysis. Table 2 suggests that economists have in general less regard for interdisciplinarity than their social scientific and

[3] The data comes from ongoing research on social science. For preliminary results, see Ollion and Abbott (forthcoming).

[4] Though economists sometimes also repurpose the techniques of others, as illustrated by the borrowing of network analysis from sociology.

Table 2

Agreement or Disagreement with the Proposition: "In general, interdisciplinary knowledge is better than knowledge obtained by a single discipline."

American university professors in	*% Agree/ Strongly agree*	*% Disagree/ Strongly disagree*	*% No answer/ Don't know*
Economics	42.1	57.3	0.6
Sociology	72.9	25.3	1.8
Political science	59.8	28.0	12.2
Psychology	78.7	9.4	11.9
Finance	86.6	9.6	3.8
History	68.2	31.7	0.1

Source: From Gross and Simmons' survey about the politics of the American professoriate. The survey was conducted in 2006. The authors sampled 100 individuals in each field. Return rates are low (though not unusually low for this kind of survey) and varied importantly across disciplines (economists: 44%; sociologists: 55%; political scientists: 54%; psychologists 49%; finance professors: 37%; historians: 54%). We are grateful to Neil Gross for running the cross-tabulations on this survey for us here and elsewhere in the paper. See Gross and Simmons (2007) for details about the survey and Gross (2013) for a broader analysis.

even business school brethren. Economists are the only ones in this group among whom a (substantial) majority disagree or strongly disagree with the proposition that "in general, interdisciplinary knowledge is better than knowledge obtained from a single discipline." Such results are consistent with the notion that economists, with their distinctive confidence in the superiority of their own discipline, are less likely to feel the need to rely on other disciplines or even to acknowledge their existence.

As sociologists know well, this dynamic is characteristic of unequal situations: those in a central position within a field fail to notice peripheral actors and are also largely unaware of the principles that underpin their own domination (Bourdieu 1984). Instead they tend to rationalize power and inequality as a "just" product of merit, justified by effort or talent. A good example of this kind of rationalization would be citing higher average scores on the Graduate Record Exam for graduate students in economics, or the higher impact factors of economics journals. Sociologists, however, might point out that such differences between fields are strongly structured by social determinants such as class, gender, and race. Meanwhile, peripheral actors compulsively orient themselves toward dominant ones, whether positively or negatively.[5]

[5] As another example of this general phenomenon, Fourcade (2006) notes that non-US-based scholars are much more likely to define their identities around the recognition they receive (or fail to receive) from American academic institutions than the other way around.

Hierarchy Within

The intellectual structure within the discipline of economics is often evoked to explain these asymmetric relations: because economists have managed to preserve a more unitary disciplinary core than other social science fields, other fields will find it easier to refer to economics, if only to establish a counterargument, than the other way around. In other words, the arguments of a unitary discipline are clearly identifiable from the outside, while those of a fractious discipline are more uncertain. Table 1 showed that citations in the *American Economic Review* are both less interdisciplinary and more concentrated than citations in the political science and sociology flagship journals. This suggests that economics more than the other fields looks both inward and toward the top of its internal hierarchy. This pattern may be interpreted in two ways: there is more consensus in economics than in sociology or political science; and there is more control. Of course these two interpretations are not mutually exclusive: there might be more consensus because there is more control (for instance if a consistent view of what constitutes quality research is promoted by those who control the top journals); conversely, control might be more effective and enforceable because there is more consensus.

There is substantial evidence that notwithstanding deep political differences amongst themselves, economists are more likely to think in a strongly integrated and unified framework than other social scientists. For instance, economists agree widely on the core set of principles and tools that structure PhD training. They also rely on textbooks much more than the other social sciences do, including at the graduate level—and graduate textbooks tend to be written by faculty from elite departments. In a survey conducted in 1990, graduate education was found to be "amazingly similar" across economics PhD programs (Hansen 1991, p. 1085).

In the interdisciplinary fellowship attribution panels studied by Lamont (2009), economists had more homogeneous standards of evaluation within, greater confidence in their judgment about research excellence even in other fields, and a higher likelihood to stick together as a group than panelists from other disciplines.[6] Only historians were similar to economists in the consistency and cohesiveness of their judgments about good historical craftsmanship, but even they were more divided internally along political lines, as well as more open to considering a variety of criteria when judging other disciplines. Judgments about the scholarly merit of proposals were more dispersed and less consensual in the humanities and other social scientific fields, making it harder to identify important works both within and without.

On the control side, economists manage their field tightly. Scholars have long noted that top departments in economics exert a remarkably strong influence over the discipline's internal labor market (Cole 1983; Whitley 1984). The most convincing

[6] Studying how mainstream economists established their position within the interdisciplinary School for Advanced Studies in the Social Sciences (EHESS) in Paris, Godechot (2011) finds a similar pattern of strong cohesion within and asymmetric relations and exclusion without.

empirical study on this point comes from the comparison by Han (2003) of the hiring process in seven disciplines (their "tribal regimes"): two from the humanities—history and English; four from the social sciences—economics, political science, psychology, and sociology; and mathematics. Using *Lingua Franca*'s annual compilations in *Job Tracks: Who Got Hired Where* (1993–2000), Han found, unsurprisingly, that all of the disciplines follow a "prestige principle": hires are strongly dependent on the prestige of departments as reported by sources such as the National Research Council and *US News and World Report.* The flows of students between departments provide unequivocal evidence: they show that universities only hire from institutions that are like-ranked or higher-ranked. Academia hence resembles the kinship systems once described by Claude Lévi-Strauss ([1949]1969), in which some alliances (between students and departments) are preferred while others, being taboo, simply can not exist. This correlation between prestige and placement, however, is strongest in economics. There, the distinctions between clusters are more clear-cut than in any other discipline. Economics departments at the very top of the pecking order exchange students amongst themselves in higher proportions than in other fields, including mathematics. Three conclusions emerge. First, hierarchy is much more clearly defined in economics. Second, the field of economics is horizontally more integrated, with strong norms of reciprocity and cohesion in recruitment processes. Third, these norms sustain a high stability of interdepartmental prestige hierarchies over time. By contrast, psychology and sociology are the most decentralized, least cohesive fields and have the least stable prestige rankings.

Getting a Job

Not simply the outcome, but also the conduct of the annual junior job market confirms these differences across the social sciences. In economics, the process is very organized, with most departments collectively deciding on the rank ordering of their own students applying for positions. This procedure, which is uncommon in many academic fields, is possible only in the context of economists' strong internal agreement on quality criteria and because of the field's belief that search and placement processes can be more efficient that way, without altering outcomes. Once a department's own students have been ranked, market intermediaries ("placement officers") are delegated with the task of helping to make matches, by proactively selling the products on offer (so to speak) to potential buyers at the other end. Finally, a ritualized evaluation process progressively filters the vetted candidates, starting with interviews at the annual meetings of the Allied Social Science Associations held in early January. For the aspiring PhD graduate, the real action at the ASSA conference takes place in the hotel suites where the hiring parties—other academic departments, but also government agencies, international institutions, and private sector firms—interview job candidates for several days on end. Meanwhile, in the public meeting rooms, the more-established scholars present their papers to their peers.

The sociology junior labor market stands in sharp contrast to this careful orchestration of the circulation of students. To job applicants and faculty in sociology, the very notion of a collectively managed process of matching students to job positions would be both unworkable in practice and objectionable in principle. To be sure, social networks play a role and informal contacts sometimes precede on-site “fly outs,” but they rarely take the form of a formal interview by a full committee, as they do in economics. Hierarchies between sociology departments are also more uncertain. A vertical structure does exist—sociologists, too, have “market stars” and keep a close eye on commonly referenced departmental rankings. But one would be hard-pressed to define the principles that underpin the pecking order in sociology. Devoid of consensual criteria for generating a putative hierarchy, and perhaps also less trusting of their colleagues’ judgment, sociologists must keep the process more open in order to build up consensus from below, inclusively. In economics, consensus is much stronger from the start; “information” about candidates is deemed homogeneous and therefore inherently reliable. As a result, the range of possible options is more tightly defined and determined much earlier.

Getting Published

The economics publications market is also comparatively more concentrated than in other social science disciplines in the sense that the most-cited journals exhibit a heavier concentration of papers coming from elite departments in economics than in sociology. This is true both in terms of the departments where authors work and the departments from which those authors graduated. For instance, according to our calculations, the top five sociology departments account for 22.3 percent of all authors published in the *American Journal of Sociology*, but the top five economics departments account for 28.7 percent of all authors in the *Journal of Political Economy* (*JPE*) and 37.5 percent in the *Quarterly Journal of Economics* (*QJE*). The contrast is even starker when one turns to the institutions from which the authors got their PhDs, with the top five sociology departments now totaling 35.4 percent in the *American Journal of Sociology*, but 45.4 percent in the *Journal of Political Econom*y and a sky-high 57.6 percent in the *Quarterly Journal of Economics*.

An economist might tend to regard this concentration as evidence that, across economics departments, intellectual strength is more concentrated in the top departments than is the case across sociology departments. Others might highlight alternative metrics that are also used for evaluation (books may be more important in some disciplines) and the existence of multiple criteria of worth, which are only imperfectly reflected in the hierarchy of scholarly journals.[7] Economists, by contrast, tend to see institutionalized hierarchies as emergent, truthful indicators of some underlying worth, and consequently are obsessed with them. For instance, in no other social science can one find the extraordinary volume of data and research about rankings (of journals, departments, and individuals) that economists

[7] On the role of books in academic careers for sociologists, see Clemens, Powell, McIlwaine, and Okamoto (1995). While the data used in this study are now 20 years old, there is no evidence that the two-pronged situation has changed much.

produce—not to mention the centralization of economic research in RePEc (an international research archive) and the continued existence of a substantial, if marginalized, subfield focused on the history of economics.

This intense awareness of hierarchies in economics breeds a fierce competition for individual status, which may explain some of the most unsettling aspects of the field's operating procedures. One notable fact is that several leading economic journals edited at particular universities have a demonstrable preference for in-house authors, while the *American Economic Review* is much more balanced in its allocation of journal space. Looking at home bias figures since the 1950s, Coupé (2004, p. 27) finds a consistent pattern of over-representation of in-house authors over time. Between 1990 and 2000 for instance, the Harvard-based *Quarterly Journal of Economics* "assigned 13.4% of its space to its own people" and 10.7 percent to neighboring MIT (against 8.8 percent to the next most prominent department, Chicago). Conversely, 9.4 percent of the pages of the Chicago-based *Journal of Political Economy* went to Chicago-affiliated scholars. This was equivalent to the share of Harvard and MIT *combined* (4.5 and 5.1 percent, respectively). Wu (2007) shows that these biases actually increased between 2000 and 2003.[8] Our data (2003–2012) confirms this domination of Cambridge, Massachusetts, over the *Quarterly Journal of Economics* and (to a lesser extent) Chicago over the *Journal of Political Economy*. The supremacy of Cambridge is even more striking when one looks at where the authors obtained their PhDs. In 2003–2012, the proportion of Harvard *graduates* publishing in the *QJE* was 20.5 percent, just edging MIT graduates (16.4 percent). Both were way ahead of the third contributor, Princeton (7.4 percent). In the *JPE*, Harvard, MIT, and Chicago graduates all hover around 10–11 percent of the authors pool.

To be sure, there are many reasons for home biases in economic journals, such as higher levels of submissions from faculty and graduate (or former graduate) students if the journal is edited in-house; a higher likelihood of being encouraged by the editor, part of whose job is to bring in good papers through interpersonal connections (Laband and Piette 1994; Medoff 2003); or journal philosophical style leading to self-selection biases in submission. But similar processes are also at play in other fields without producing the same dramatic effects. Thus, even if the social structure of the field may explain some of these differences, it does not explain them *away*: the structure itself stands at the core of the phenomenon that interests us here, which is the stable supremacy of three departments—Chicago, Harvard, and MIT—over the rest of the field, bolstered via control over two university-based journals. As a point of comparison, such home bias is virtually

[8] Wu (2007) finds that 14 percent of *JPE* pages published over that period went to Chicago authors, and a whopping 28 percent of *QJE* pages went to Harvard–MIT authors (specifically, 15 percent for Harvard and 13 percent for MIT). Our data for the 2003–2012 period shows that the University of Chicago still ranks first with 10.8 percent of the total authors published in the *Journal of Political Economy*, followed by Harvard (6.1 percent) and the MIT (4.1 percent). During the same period, the *Quarterly Journal of Economics* published almost twice as many authors (14.9 percent) from Harvard than from Chicago (7.0), with the MIT coming third (6.2 percent).

nonexistent in the main sociology journal edited out of a university department, the *American Journal of Sociology*, which is based at the University of Chicago.[9] This suggests that the pattern of home bias in top economics journals, together with the stability of rankings of top departments, is not just a coincidence of geography and authors, but stems instead from a particular form of social organization and control.

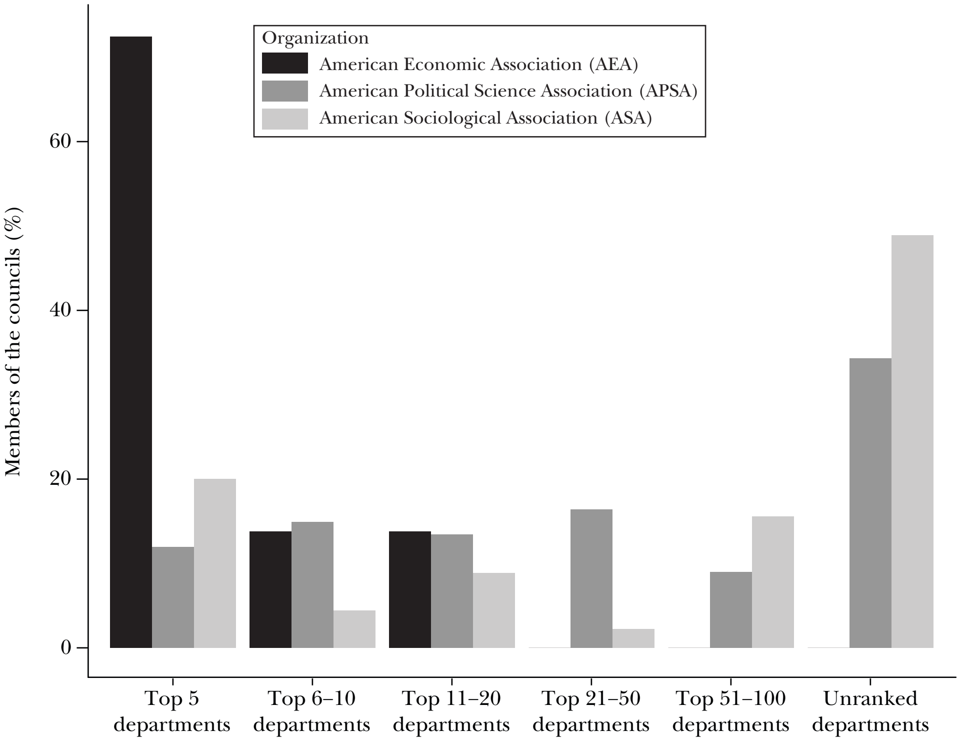
Getting Together

Finally, looking at professional associations across social scientific fields confirms the more cohesive and hierarchical organization of economics, and the more fractious character of its sister disciplines. A rapid comparison of the by-laws of the American Economic Association (AEA), the American Sociological Association (ASA), and the American Political Science Association (APSA) shows vast disparities in the distribution of political power across the disciplines. Despite being 18,000 members strong, the AEA is a minimalist organization based out of Nashville, Tennessee. Dues are low, at $20–$40/year as of 2014. The by-laws are short, at 1,770 words, and procedures are centralized. There are only six elected officers, and only one candidate typically runs for president-elect. As Figure 2 shows in dramatic fashion, the AEA leaders are drawn disproportionately from the discipline's elite departments: that is, 72 percent of the AEA nonappointed council members are from the top five departments, in contrast with only 12 and 20 percent respectively for APSA and ASA. The president-elect and program committee run the program for the annual meetings, which involves selecting ahead of time the sessions to be conducted and the papers from a subset of the sessions to be included in the "Papers and Proceedings" issue of the *American Economic Review* (the May issue following the annual meeting). This procedure ensures a flagging of topics and authors deemed most important by the organization's leadership.

This approach contrasts with the more internally balkanized and also more grassroots nature of the American Sociology Association and the American Political Science Association. Although these professional associations have fewer members than the American Economic Association (about 15,000 for APSA and 13,000 for ASA), their staffs are larger. Procedures are more complex, as reflected in the length of their by-laws: 4,657 words for the ASA, 5,529 for APSA. While the AEA is a unitary organization, community life among sociologists and political scientists revolves around "sections" or organized subfields, each of which has its own procedures, dues, awards, and program at the annual meeting. The ASA solves the political problem of internal divisions by having contested elections at both the central and section levels while the APSA has long resorted to institutionalized horse-trading between the dominant constituencies. In both cases, the organizations' leaders are

[9] If anything, our data suggest that there might instead be a bias *against* Chicago faculty in the *American Journal of Sociology*, who barely make it to the top 20 with a mere 1.4 percent of published papers. Although this proportion rises to 6.9 percent for former Chicago sociology graduates, they are still topped by both Harvard PhDs (9.4 percent) and Stanford PhDs(8 percent).

Figure 2
Institutional Composition of the Executive Council of Three Disciplinary Organizations by Ranking of Departments of the Nonappointed Officers, 2010–2014



Notes: The ranking of departments is based on the *U.S. News and World Report* 2012 ranking of best graduate schools, by discipline. The "unranked" category comprises mostly departments that do not have a graduate program, and a very small number of foreign institutions.

drawn primarily from nonelite institutions, as shown in Figure 2. Because the disciplinary core is less identifiable and more contested, members of the ASA and the APSA also *identify less* with it: the rank-and-file is less bound to the elite and both associations fulfill primarily a democratic purpose of integration across the board, an openness that is also reflected in the structuring of their conference programs. However, the marginalization of most of the association leaders at the ASA and the APSA from the high-prestige core of the discipline, and also from political power, also explains both organizations' frantic striving for influence, manifested, among other things, in their Washington addresses. To support this more elaborate infrastructure and expensive residence, dues for both organizations are among the highest in the social sciences: $50 to $350/year for the ASA; and $40 to $320 for the APSA—not counting section dues.

The Rise of Finance

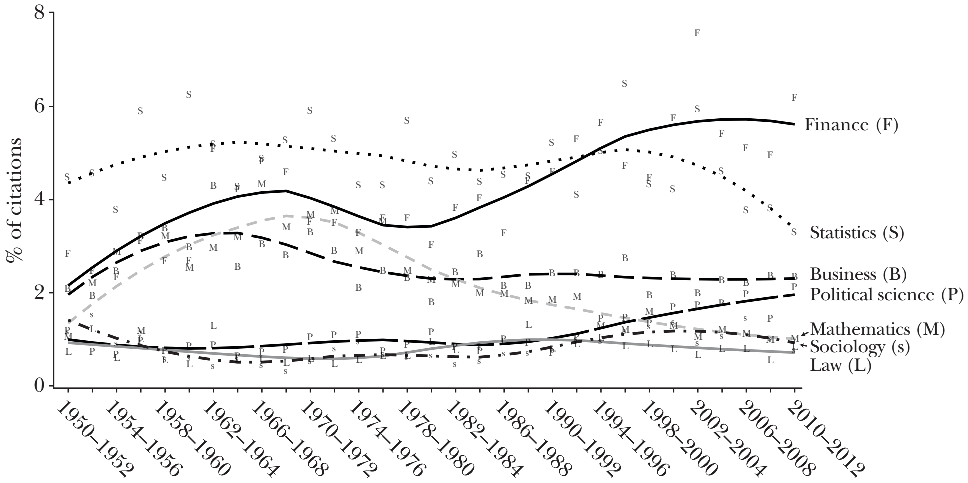
For all the relative insularity and autonomy of economics, economists still do engage other disciplines. Our analysis of five top economics journals shows that between 19 and 25 percent of citations are outside the discipline, a fairly stable pattern since the end of World War II. But when economics goes interdisciplinary, where does it turn? Have the disciplinary connections of economics changed over time, and if so, what does this tell us about the evolution of the field?

This framing provides us with a different road into the recent history of economics than much of the literature, which is often focused tightly on trends within economies: examples include the transformation of publishing patterns in economic journals (Card and DellaVigna 2013), the rise and fall of fields within economics in volume (Kelly and Bruestle 2011) and in relative prestige (Ellison 2010), or the downward trend in the use of mathematics and in the publication of theoretical papers (Hamermesh 2013). Instead, we begin by analyzing the network of relations between economics and other disciplines over time. In other words, we start from the assumption that who you cite says something about who you are. We find that changing patterns of external citations indeed tell us quite a lot about the inner situation of the discipline and the changing relative power of different constituencies.

Figure 3 offers a representation of economics' extra-disciplinary references, based on our extensive study of citations in five top economics journals that were all founded well before World War II: the *Quarterly Journal of Economics* (founded in 1899), the *Journal of Political Economy* (1899), the *American Economic Review* (1911), *Econometrica* (1933), and the *Review of Economic Studies* (1933).[10] The figure tells a story that is partly familiar, partly less so. The points in the figure show the share of outside-the-field citations in economics journals going to journals in the fields of finance (F), statistics (S), business (B), political science (P), mathematics (M), sociology (s), and law (L). Because there is considerable fluctuation from year to year, we show the patterns of the data as smoothed curves. The figure shows the dramatic rise of economics' engagement with mathematics and statistics in the post-World War II period. The high point of this engagement, in the mid-1970s, coincides with the low point of engagement with the other social sciences (such as political science and sociology), as well as with practical enterprises, such as law and, with a slight delay, business. Notwithstanding the foundations' and government's efforts to promote interdisciplinary ventures under the "behavioral sciences" label in the 1950s, the social sciences became clearly more estranged from one another in the 1960s–70s. Nor was economics the only driving force in this process: cross-disciplinary experiments at Harvard (the Department of Social Relations) and Carnegie-Mellon failed, and all the various fields retreated into their own distinctive form of abstraction and high theory (Steinmetz 2005; Isaac 2010).

[10] Citations were obtained from the Institute for Scientific Information's Web of Social Science. The lines were drawn using a smoothing coefficient. See the online Appendix available with this paper at http://e-jep.org for details.

Figure 3
Extradisciplinary Citation in Five Top Economics Journals
(to papers in fields of finance, statistics, business, political science, mathematics, sociology, and law)

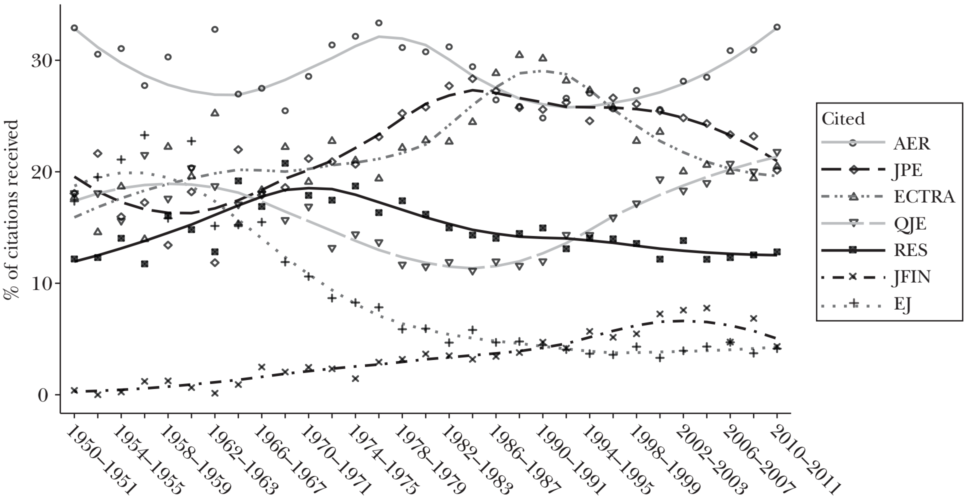


Source: The raw citation data were collected from the Institute of Scientific Information's *Web of Social Sciences.*
Notes: The points in the figure show the share of outside-the-field citations in five economics journals going to journals in the fields of Finance (F), Statistics (S), Business (B), Political Science (P), Mathematics (M), Sociology (s), and Law (L). The top five economics journals are the *Quarterly Journal of Economics* (founded in 1899), *Journal of Political Economy* (1899), *American Economic Review* (1911), *Econometrica* (1933), and *Review of Economic Studies* (1933). We show the patterns of the data as smoothed curves. The lines were drawn using a smoothing coefficient. See the online Appendix available with this paper at http://e-jep.org for details.

The interdisciplinary ecology as it stood toward the end of the period depicted in Figure 3 looks very different. Citations to mathematics in the leading economics journals are practically gone and those to statistics have faltered. The other social sciences have made a modest comeback, particularly political science (which has had a partial conversion to rational choice theory). But the most striking trend from Figure 3 in recent decades is the continuous rise of finance as a purveyor of "interdisciplinary" references for economics.

In judging the magnitude of this trend toward finance, it is important to note that our estimate of the rise of the role of finance within economics in Figure 3 is very conservative. Our list of five top economics journals does not include any finance journal. Figure 4 presents an analysis of citations among our list of five top economics journals plus two more: the *Journal of Finance* (founded in 1946) over time; and the British-based *Economic Journal* (founded in 1891)—a core generalist publication for economists for much of the twentieth century, on par with the *JPE* and *QJE* at the beginning of the period. Self-citations are not counted in the total of cross-citations. Reading the graph, we see that in 2010–2011, the AER got 33 percent

Figure 4
Citations among Six Economics Journals and One Finance Journal
(excluding self-citations; (Quarterly Journal of Economics (QJE)*,* Economic Journal (EJ)*,* Journal of Political Economy (JPE)*,* American Economic Review (AER)*,* Econometrica (ECTRA)*,* Review of Economic Studies (RES)*, and* Journal of Finance (JFIN)*)*



Source: The raw citation data were collected from the ISI's *Web of Social Sciences*.
Notes: Figure 4 presents an analysis of citations among six economics journals plus the *Journal of Finance*. Self-citations are not counted in the total of cross-citations. Reading the graph, we see that in 2010–2011, the *AER* got 33 percent of the cross-citations among that set of journals, self-citations removed. In addition to the data points, the graph shows lines drawn using a smoothing coefficient.

of the cross-citations among that set of journals, self-citations removed. The graph shows a lot of action at the top—the meteoric return of the *QJE* to prominence, the relative decline of *Econometrica* and *JPE*—but two other salient transformations over the very long run are the constant decline of the British journals (*RES* and *EJ*), particularly the *EJ* (which disappears into near-oblivion) and the rise of the *Journal of Finance*. Our bibliometric network data (not shown) indicates that by the 2000s, the *JF* was most closely integrated with the core US-based publications, receiving between 7 and 11 percent of all the cross-references (excluding self-citations) in the *AER*, *QJE* and *JPE*. In other words, the *JF*, which would not have been considered an economics journal when it was first founded, has become an integral part of the economics disciplinary matrix. Other finance journals have followed suit, too, as financial economics has become the dominant approach in the field (Jovanovic 2008).[11]

[11] The first issue of the *Journal of Financial Economics* came out in 1974, and it is now ranked as the eighth economics journal by impact factor. The *Review of Financial Studies*, first published in 1988, ranks twelfth. The *Journal of Finance* now ranks fifth by impact factor overall in economics, edging ahead of the much older *Review of Economic Studies*.

The institutional rise of finance as an intellectual powerhouse within economics follows from the establishment of a teaching base in business schools in the second half of the twentieth century. Over that period, business schools, which control the production of certified managers (through the MBA degree), have evolved from practitioner-dominated programs struggling for academic legitimacy to become the largest employers of trained social scientists, now rivaling traditional academic departments in the size and distinction of their faculties. A survey from 2004 found 549 economics PhDs teaching in the top 20 US business schools, as compared with 637 economics PhDs in the top 20 economics departments (Blau 2006). This absorption of increasingly large contingents of economics PhDs has turned business schools into formidable players within economic science itself—a transformation that is attested by the remarkable string of Nobel Prizes in economic science awarded to scholars based in business schools since 1990 (Fourcade and Khurana 2013), including Eugene Fama, Oliver Williamson, Robert Engle, Michael Spence, Robert Merton, Myron Scholes, Merton Miller, John Harsanyi, and Robert Fogel.

Our own analysis of papers published in the *American Economic Review* since the 1950s reveals a rapid rise in business school affiliations among authors, and a simultaneous and sharp decline in government-based authors. The share of authors whose primary affiliation is to a business school has increased steadily from a low 3.2 percent in the 1950s to 17.9 percent in the 2000s. Conversely, contributions from scholars located in government agencies have become marginal.[12]

As the academic field of economics shifted toward business schools—and away from government—economists faced a new set of practical, intellectual, and political entanglements: higher levels of compensation, new connections and consulting opportunities, and often different politics as well (Jelveh, Kogut, and Naidu 2014). In the 1980s, suspicion of government action grew markedly within the field, and economists arguably supplied part of the intellectual rationale for the deregulatory movement in public policy and for the expanded use of price and market mechanisms in education, transportation, healthcare, the environment, and elsewhere (Blyth 2002). Financial economists argued forcefully that the purpose of corporations was to maximize shareholder value, and provided a scientific justification for the management practices favored by a new generation of corporate raiders: leveraged buy-outs, mergers and acquisitions, and compensating corporate executives with stock options.[13] In a recent indictment of the “pervasiveness of the capture of economists by business interests,” Zingales (2013) found that, when none of their authors worked in a business school, economics articles were significantly

[12] Measures are based on self-declared affiliations on the articles we surveyed. When authors mentioned several affiliations (a trait that has increased over time), we adopted the following procedure: If there was a clear order, we opted for the first institution. Otherwise, and in an attempt to not artificially increase the share of secondary affiliations, we gave priority to “economics department” when mentioned equally with any another institution. See the online Appendix available with this paper at http://e-jep.org for details.

[13] For instance, see Fligstein and Shin (2007), Jung and Dobbin (2012), Fourcade and Khurana (2013), and Heilbron, Verheul, and Quak (2014).

"less likely to be positive on the level of executive compensation, and significantly more likely to be negative" (p. 139).[14]

A Life of Their Own

Economists have distinctive opinions, beliefs, and tastes compared to academics in other fields and to the broader American public. Evidence on this topic is dispersed and must be pieced together from various sources. A sizeable share comes from economists themselves: the home-grown literature on the topic is abundant. The field is filled with anxious introspection, prompted by economists' feeling that they are powerful but unloved, and by robust empirical evidence that they are different. In some classic examples, Marwell and Ames (1981) found that first-year graduate students in economics at the University of Wisconsin were less likely to make contributions to a public good in a structured laboratory game. In this journal, Frank, Gilovich, and Regan (1993) cite a range of evidence suggesting that studying economics inhibits cooperation. The extent to which such differences persist across the contexts of different laboratory studies, and the underlying cause of any differences that do persist, remains controversial. Is it that learning economics makes people more accepting of self-interested behavior in themselves and others? Or perhaps it is that the discipline attracts more egoistic people? Frey and Meier (2005) look at voluntary student contributions to social funds at the University of Zurich, and find that those who will later choose economics as a field of study are less likely to contribute—even before their economic studies begin. Whatever the underlying dynamic, there is suggestive and convergent evidence that economists are either more candid about pursuing their self-interest, or simply more selfish (by disposition or as a result of training).

Economists are likely to find themselves in a minority position on some of their dearest ideas. Sapienza and Zingales (2013) argue that the more American economists agree among themselves, the more distant they grow from average Americans. In general, of course, economists favor using market-based solutions to address social issues (Whaples 2009). They support allowing payments to be made to organ donors, but the public finds the very thought distasteful. A sizeable majority of economists believes that trade protectionism is economically harmful, but when asked whether "buying American" is good for the economy, the average American agrees it is (Sapienza and Zingales 2013, p. 638). Economists think that a market mechanism such as a carbon tax or a cap-and-trade system of pollution permits is a more cost-effective mechanism to curb climate change than regulatory steps such as car emissions standards, but most of their fellow citizens beg to differ. Economists may advise governments, but they often do not convince the people.

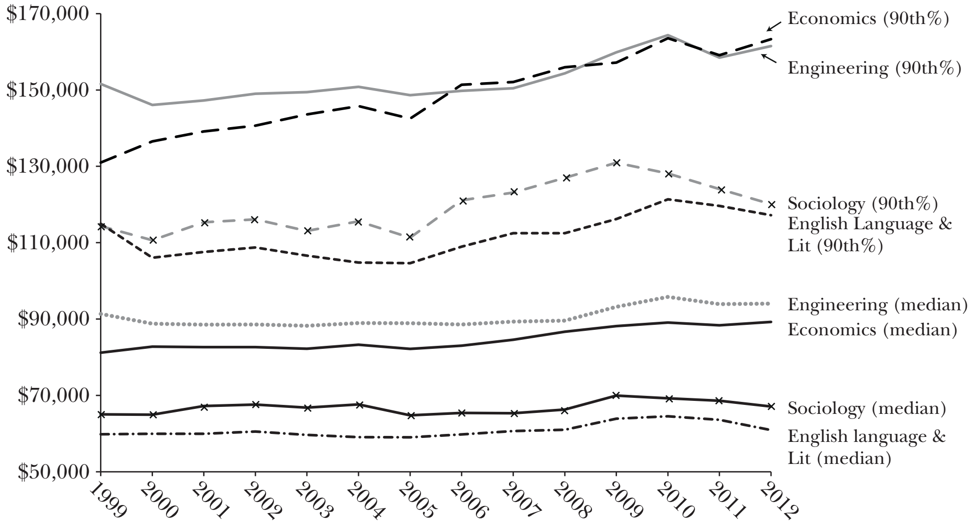
[14] The sample included 150 of the most cited downloaded SSRN papers prior to 2008 using the search key word "executive compensation" (excluding survey papers).

Academic economists vote more to the left than American citizens, like most of their university-based peers. They have been doing so for as long as political opinion surveys have been administered in this setting: Ladd and Lipset (1976) offer a classic early survey. Even though on average the contingent of libertarians among economists is much larger than among the US voting public, as a group, economists still claim to trust the government more—with some important institutional variations. According to the Gross and Simmons survey of the American professoriate (see Gross 2013), economists are situated about halfway between humanities scholars and other social scientists to their left and business schools professors to their right in most of their political opinions. For example, two-thirds of sociologists say that corporations make too much profit, but only one-third of economists and virtually no finance professors think so. The overwhelming majority of sociologists (90 percent) endorse the proposition that "the government should do more to help needy Americans, even if it means going deeper into debt," but barely one-half of the economists and one-third of the finance scholars agree with that proposition.

The worldviews of economists, like those of all individuals, are in part the product of their particular social entanglements—the material and symbolic situation and trajectory of their group, and that of each individual within it. Relative to other academics, economists do better in terms of income. According to the Bureau of Labor Statistics, the mean salary for the 11,000 economics teachers in colleges, universities, and professional schools was $103,000 in 2012, and $160,000 for the top 10 percent. For comparisons, the mean figure for sociologists was $76,000, with the top 10 percent at $118,000. These totals do not count additional sources of income from consulting work or other activities, which can be substantial (Weyl forthcoming). Furthermore, economists' material situation has improved noticeably over the last two decades, particularly for the best-paid members of the profession, who now narrowly outstrip the best-paid engineers; by contrast, the median real wage in many academic professions (the humanities, mainly) and in the United States at large barely rose over the same period, as we see in Figure 5 (which also covers teachers at junior colleges in addition to colleges, universities, and professional schools). How this experience of group social mobility and growing intrafield inequalities may have affected economists' appreciation of the deteriorating relative economic situation of their less-fortunate fellow academics and citizens is an open question.

This growing social distance of economists from the public at large would be irrelevant if economists were not making it their mission to maximize the welfare of ordinary people. Economics as a profession is prominently intertwined with public administrations, corporations, and international organizations; these institutions not only provide economists with resources and collect their data, they also foster a "fix it" culture—or, as sociologists would put it, a particular "habitus," a disposition to intervene in the world (Bourdieu and Wacquant 1992). Economists, particularly modern economists, want to fix things, which is both a product of their theoretical confidence and of the position of their discipline within society (Mitchell 1998). For instance, economic models routinely invoke the mythical figure of the benevolent

Figure 5
Annual Median and 90th Percentile Wages in Selected Disciplines, 1999–2012
(2012 constant dollars)



Source: Bureau of Labor Statistics (BLS) Occupational Employment Statistics, http://www.bls.gov/oes/tables.htm.
Note: Figure 5 covers teachers at junior colleges in addition to colleges, universities, and professional schools.

"social planner," imagining what this entity would do to make the world richer, healthier, and less vulnerable to shocks. Economists have developed a precise theoretical framework for evaluating when markets produce efficiency and when market failures can occur, and they have a vast econometric arsenal at their disposal to parse out the effects of actual policy proposals. In the last quarter of the twentieth century, they also started running narrowly specified field experiments, increasingly putting the administration of social policy or development aid at the service of research (for example, Banerjee and Duflo 2013). (One may note in passing that the experiments of economists are quite different from those of sociologists, who tend to run experiments to understand how people *live*.) Finally, economists are fairly certain about their ultimate judgment criteria—their predilection for efficiency over fairness, the eliciting of preferences from behavior, and the design of experiments around a tight menu of choices. These criteria positively sanction both an orientation toward policy adjudication and advice, and a distinctive willingness, even eagerness to serve and intervene. If things don't work the way they *should*, then a smart readjustment, a "nudge," may even be called for (Thaler and Sunstein 2008).

Here again, a comparison with sociologists is telling: sociologists might vie for the position of the prince's counselor too, but they have been much less successful at securing influence. First, economics and sociology have different orientations to time. Economists generally pay little attention to history, "live in the now," and "see trajectories from the present forward," while sociologists have the reverse intellectual attitude, looking at the present as the outcome of a set of past processes (Abbott 2005). Thus, sociologists often find themselves both effectively marginalized *and* shying away from direct policy involvement. Their intellectual habits center around social critique precisely because they are already outside: in the words of sociologist Pierre Bourdieu, they "make a virtue of necessity." Self-perceptions reflect these differences well. In Gross and Simmons's (2007) survey of the American professoriate, economists described themselves mainly as "intellectuals" and "scientists." Sociologists were most comfortable with the terms "social critics" and "scientists," unconsciously embracing their own peripheral position but without abandoning the mantel of science. The combination of sociologists' desire for relevance with their deep ambivalence toward power produces a very different set of dispositions: sociologists analyze critically, sometimes rouse and stir, but they rarely venture to propose fixes and remedies (they are not in a position to do so and would perhaps be reluctant to even if they had the opportunity). Political scientists, interestingly, saw themselves primarily as "intellectuals," but perhaps reflecting their much closer proximity to the political game, they were also somewhat more likely to distance themselves from the label "scientists" than either sociologists or economists.

The upshot of economists' confident attitude toward their own interventions in the world is that economics, unlike sociology or political science, has become a powerful transformative force. Economists do not simply depict a reality out there, they also make it happen by disseminating their advice and tools. In sociological terms, they "perform" reality (Callon 1998). Aspects of economic theories and techniques become embedded in real-life economic processes, and become part of the equipment that economic actors and ordinary citizens use in their day-to-day economic interactions. In some cases, the practical use of economic technologies may actually align people's behavior with its depiction by economic models. By changing the nature of economic processes from within, economics then has the power to make economic theories truer. For example, MacKenzie (2006) discusses how academic financial theories gave rise to enormous markets in futures, options, and other derivative financial instruments: the use of the Black–Scholes–Merton formula by market actors altered economic processes in such a way that it improved the fit of the model to the reality of option prices.

The world has changed in important ways under the influence of economists. Economic reasoning, expertise, and technologies permeate capitalist activities, culture (including the media and best-seller lists), and institutions, from hospitals through courts to universities (Hirschman and Popp Berman 2014). Economists dispense their expertise on practically all matters of public policy and have made steady gains in business and government, often in top political positions (Montecinos and Markoff 2009). Finance ministries, central banks, government agencies, international organizations, and dominant consultancies harbor large concentrations of professionally trained economists, who claim tutelary power over "the economy" while viewing societies as involved in a never-ending but ultimately beneficial process of economic reconstruction. Finally, the rational-formalist language of the economics profession underpins its universalistic aspirations. Economic fashions circulate across borders, drawing people and techniques in their wake. Much more than sociology or political science, economics is a symbolically and materially globalized discipline (Fourcade 2006).

Thus, most economists feel quite secure about their value-added. They are comforted in this feeling by the fairly unified disciplinary framework behind them, higher salaries that many of them believe reflect some true fundamental value, and a whole institutional structure—from newspapers to congressional committees to international policy circles—looking up to them for answers, especially in hard times. In fact, the recent economic and financial crisis has arguably made the discipline of economics as a whole more, not less, visible, and its expertise more sought-after: the deep recessions of the early 1980s and the Great Depression of the 1930s had the same effect.

But because economics is a transformative force, and because its operatives tend to be in charge, economists are also more exposed. The financial and economic maelstrom of 2008, which few in the economics profession had anticipated (but whose institutional roots could be traced back, in part, to actions some of them had lobbied for), led many economists to engage in soul-searching about their lack of awareness, their intellectual bullishness, and the reliability of their claims to expertise. Following discomforting interviews in the 2010 movie documentary *Inside Job*, in which prominent members of the profession emphatically denied the possibility of conflicts of interest for economists, the American Economic Association promoted a set of ethical guidelines. From his powerful tribune at the *New York Times*, Nobel-prize winner Paul Krugman (2009) aired the dirty laundry of macroeconomics—usually buried in esoteric models—in a fierce and very public manner. Economists also began to talk about distributional issues, the bread and butter of that other social science, sociology, in a way that was unimaginable just two or three decades before. To be sure, the changing facts of inequality warrant this newfound interest (Piketty 2014). But the intellectual winds in economics may be shifting, too.

Conclusion: Humble, Competent People?

"If economists could manage to get themselves thought of as humble, competent people, on a level with dentists, that would be splendid!" Keynes ([1931] 1962, p. 373) famously wrote. Most modern economists have a strong practical bent. They believe in the ideal of an expert-advised democracy, in which their competence would be utilized and on display in high-profile, non-elective positions in government and other institutions. But democratic societies are deeply suspicious of (nondemocratic) expertise; and economic advice, unlike dentistry, can never be humble. The fact is that—in some ways true to its philosophical origins—economics is a very moral science after all. Unlike atoms and molecules, the "objects" upon which economists seek to act have a perspective on the world, too. Human life is messy, never to be grasped in its full complexity or shaped according to plan: people act in unanticipated ways; politics makes its own demands; cultures (which economists do not understand well) resist. Thus, the very real success of economists in establishing their professional dominion also inevitably throws them into the rough and tumble of democratic politics and into a hazardous intimacy with economic, political, and administrative power. It takes a lot of self-confidence to put forward decisive expert claims in that context. That confidence is perhaps the greatest achievement of the economics profession—but it is also its most vulnerable trait, its Achilles' heel.

■ *The authors would like to warmly thank Johannes Buggle, Alizée Delpierre, Fatine Guedira, Daniel Kluttz, and Maxime Sauzet for research assistance. We are grateful to David Autor, Julien Boelaert, Claude Fischer, Eric Giannella, Arthur Goldhammer, Peter Gourevitch, Pierre-Olivier Gourinchas, Johan Heilbron, Chang-Tai Hsieh, Daniel Kluttz, Ulrike Malmendier, Philippe Martin, Victor Nee, Gérard Roland, Heddy Riss, Timothy Taylor, and the participants of the CCOP seminar at UC Berkeley for helpful comments.*

References

Abbott, Andrew. 2001. *Chaos of Disciplines.* University of Chicago Press.

Abbott, Andrew. 2005. "The Idea of Outcome in U.S. Sociology." In *The Politics of Methods in the Human Sciences: Positivism and Its Epistemological Others*, edited by George Steinmetz, 393–426. Durham, NC: Duke University Press.

Angrist, Joshua D., and Jórn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2): 3–30.

Banerjee, Abhijit V., and Esther Duflo. 2013. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. Random House.

Baron, James N., and Michael T. Hannan. 1994. "The Impact of Economics on Contemporary Sociology." *Journal of Economic Literature* 32(3): 1111–46.

Blau, Francine D. 2006. "Report of the Committee on the Status of Women in the Economics Profession." *American Economic Review* 96(2): 519–26.

Blyth, Mark. 2002. *Great Transformations: Economic Ideas and Institutional Change in the Twentieth Century.* Cambridge University Press.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgment of Taste.* Cambridge: Harvard University Press.

Bourdieu, Pierre. 1988. *Homo Academicus.* Stanford University Press.

Bourdieu, Pierre, and Loïc J. D. Wacquant. 1992. *An Invitation to Reflexive Sociology.* University of Chicago Press.

Bowles, Samuel. 1998. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions." *Journal of Economic Literature* 36(1): 75–111.

Callon, Michel. 1998. "Introduction: The Embeddedness of Economic Markets in Economics." In *The Laws of the Markets,* 1–57. Oxford: Blackwell.

Card, David, and Stefano DellaVigna. 2013. "Nine Facts about Top Journals in Economics." *Journal of Economic Literature* 51(1): 144–61.

Clemens, Elisabeth S., Walter W. Powell, Kris McIlwaine, and Dina Okamoto. 1995. "Careers in Print: Books, Journals, and Scholarly Reputations." *American Journal of Sociology* 101(2): 433–94.

Colander, David. 2005. "The Making of an Economist Redux." *Journal of Economic Perspectives* 19(1): 175–98.

Cole, Stephen. 1983. "The Hierarchy of the Sciences?" *American Journal of Sociology* 89(1): 111–39.

Coupé, Tom. 2004. "Revealed Performances. Worldwide Rankings of Economists and Economics Departments 1969–2000." Unpublished paper (updated version), Université Libre de Bruxelles, ECARES. Available at http://web.archive.org/web/20070717035652/http://homepages.ulb.ac.be/~tcoupe/ranking.html.

Ellison, Glenn. 2010. "How Does the Market Use Citation Data? The Hirsch Index in Economics." NBER Working Paper 16419.

Ellison, Glenn. 2011. "Is Peer Review in Decline?" *Economic Inquiry* 49(3): 635–57.

Fehr, Ernst, and Karla Hoff. 2011. "Introduction: Tastes, Castes and Culture: The Influence of Society on Preferences." *Economic Journal* 121(556): F396–F412.

Fligstein, Neil, and Taekjin Shin. 2007. "Shareholder Value and the Transformation of the U.S. Economy, 1984–2000." *Sociological Forum* 22(4): 399–424.

Fourcade, Marion. 2006. "The Construction of a Global Profession: The Transnationalization of Economics." *American Journal of Sociology* 112(1): 145–94.

Fourcade, Marion. 2009. *Economists and Societies: Discipline and Profession in the United States, Great Britain, and France, 1890s to 1990s.* Princeton University Press.

Fourcade, Marion, and Rakesh Khurana. 2013. "From Social Control to Financial Economics: the Linked Ecologies of Economics and Business in Twentieth Century America." *Theory and Society* 42(2): 121–59.

Frank, David J., and Jay Gabler. 2006. *Reconstructing the University: Worldwide Shifts in Academia in the 20th Century.* Stanford University Press.

Frank, Robert H., Thomas Gilovich, and Dennis T. Regan. 1993. "Does Studying Economics Inhibit Cooperation?" *Journal of Economic Perspectives* 7(2): 159–71.

Freeman, Richard B. 1999. "It's Better Being an Economist (But Don't Tell Anyone)." *Journal of Economic Perspectives* 13(3): 139–45.

Frey, Bruno S., and Stephan Meier. 2005. "Selfish and Indoctrinated Economists?" *European Journal of Law and Economics* 19(2): 165–171.

Godechot, Olivier. 2011. "How Did the Neoclassical Paradigm Conquer a Multi-disciplinary Research Institution?" *Revue de la régulation* 10(Autumn). http://regulation.revues.org/9429.

Gross, Neil. 2013. *Why Are Professors Liberal and Why Do Conservatives Care?* Cambridge, MA: Harvard University Press.

Gross, Neil, and Solon Simmons. 2007. "The Social and Political Views of American Professors." Working paper accessed March 10, 2014: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.6141&rep=rep1&type=pdf.

Hamermesh, Daniel S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature* 51(1): 162–72.

Han, Shin-Kap. 2003. "Tribal Regimes in Academia: A Comparative Analysis of Market Structure across Disciplines." *Social Networks* 25(3): 251–80.

Hansen, W. Lee. 1991. "The Education and Training of Economics Doctorates: Major Findings of the Executive Secretary of the American Economic Association's Commission on Graduate Education in Economics." *Journal of Economic Literature* 29(3): 1054–87.

Haskell, Thomas L. 2000. *The Emergence of Professional Social Science: The American Social Science Association and the Nineteenth-Century Crisis of Authority.* Johns Hopkins University Press.

Heilbron, Johan, Jochem Verhael, and Sander Quak. 2014. "The Origins and Early Diffusion of 'Shareholder Value' in the United States." *Theory and Society* 43(1): 1–22.

Hirshman, Daniel, and Elizabeth Popp Berman. 2014. "Do Economists Make Policies? On the Political Effects of Economics." *Socio-Economic Review* 12(4): 779–811.

Isaac, Joel. 2010. "Theorist at Work: Talcott Parsons and the Carnegie Project on Theory, 1949–1951." *Journal of the History of Ideas* 71(2): 287–311.

Jacobs, Jerry A. 2013. *In Defense of Disciplines: Interdisciplinarity and Specialization in the Research University.* University of Chicago Press.

Jelveh, Zubin, Bruce Kogut, and Suresh Naidu. 2014. "Political Language in Economics." Unpublished paper.

Jovanovic, Franck. 2008. "The Construction of the Canonical History of Financial Economics." *History of Political Economy* 40(2): 213–42.

Jung, Jiwook, and Frank Dobbin. 2012. "Finance and Institutional Investors." In *The Oxford Handbook of the Sociology of Finance*, edited by Karin Knorr Cetina and Alex Preda, 52–74. Oxford University Press.

Kelly, Michael A., and Stephen Bruestle. 2011. "Trends of Subjects Published in Economic Journals, 1969–2007." *Economic Inquiry* 49(3): 658–73.

Keynes, John Maynard. [1931]1962. *Essays in Persuasion.* W. W. Norton.

Krugman, Paul. 2009. "How Did Economists Get It So Wrong?" *New York Times Magazine.* September 2.

Laband, David N., and Michael J. Piette. 1994. "Favoritism versus Search for Good Papers: Empirical Evidence Regarding the Behavior of Journal Editors." *Journal of Political Economy* 102(1): 194–203.

Ladd, Everett Carll, Seymour Martin Lipset. 1976. *The Divided Academy: Professors and Politics*. New York: W. W. Norton & Company.

Lamont, Michéle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment*. Harvard University Press.

Lazear, Edward P. 2000. "Economic Imperialism." *Quarterly Journal of Economics* 115(1): 99–146.

Leamer, E. 2010. "Tantalus on the Road to Aymptotia." *Journal of Economic Perspectives* 24(2): 31–46.

Lebaron, Frédéric. 2000. *La Croyance économique. Les économistes entre science et politique*. Paris: Seuil

Lévi-Strauss, Claude. [1949] 1969. *The Elementary Structures of Kinship*. Boston, MA: Beacon Press.

MacKenzie, Donald. 2006. *An Engine, Not a Camera: How Financial Models Shape Markets*. Cambridge, MA: MIT Press.

Marwell, Gerald, and Ruth E. Ames. 1981. "Economists Free Ride, Does Anyone Else? Experiments on the Provision of Public Goods." *Journal of Public Economics* 15(3): 295–310.

Medoff, Marshal H. 2003. "Editorial Favoritism in Economics?" *Southern Economic Journal* 70(2): 425–34.

Mirowski, Philip. 1989. *More Heat than Light. Economics as Social Physics, Physics as Nature's Economics*. Cambridge University Press.

Mitchell, Timothy. 1998. "Fixing the Economy." *Cultural Studies* 12(1): 82–101.

Montecinos, Verónica, and John Markoff, eds. 2009. *Economists in the Americas*. Edward Elgar.

Nik-Khah, Edward, and Robert Van Horn. 2012. "Inland Empire: Economics' Imperialism as an Imperative of Chicago Neoliberalism." *Journal of Economic Methodology* 19(3): 259–82.

Ollion, Étienne, and Andrew Abbott. Forthcoming. "Quarante ans de sociologie française aux États-Unis. Note sur la réception des sociologues français Outre-Atlantique (1960–2009)." In *Transmissions. Une communauté en héritage*, edited by D. Demazière, D. Lorrain, C. Paradeise. Paris: PUR.

Pieters, Rik, and Hans Baumgartner. 2002. "Who Talks to Whom? Intra- and Interdisciplinary Communications of Economic Journals." *Journal of Economic Literature* 40(2): 483–509.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press.

Prasad, Monica. 2006. *The Politics of Free Markets: The Rise of Neoliberal Economic Policies in Britain, France, Germany, and the United States*. University of Chicago Press.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ross, Dorothy. 1991. *The Origins of American Social Science*. Cambridge University Press.

Sapienza, Paola, and Luigi Zingales, L. 2013. "Economic Experts versus Average Americans." *American Economic Review* 103(3): 636–42.

Sims, Christopher A. 2010. "But Economics is not an Experimental Science." *Journal of Economic Perspectives* 24(2): 59–68.

Steinmetz, George, ed. 2005. *The Politics of Method in the Human Sciences: Positivism and Its Epistemological Others*. Duke University Press.

Stigler, George J., and Gary Becker. 1977. "De Gustibus Non Est Disputandum." *American Economic Review* 67(2): 76–90.

Richard H. Thaler, and Cass R. Sunstein. 2008. *Nudge. Improving Decisions about Health, Wealth and Happiness*. Penguin Books.

Weyl, Glen. Forthcoming. "Finance and the Common Good." Conclusion of *Après le Déluge: Finance and the Common Good after the Crisis*, edited by Edward Glaeser, Tano Santos, and Glen Weyl. University of Chicago Press.

Whaples, Robert. 2009. "The Policy Views of American Economic Association Members: The Results of a New Survey." *Economic Journal Watch* 6(3): 337–48.

Whitley, Richard. 1984. *The Intellectual and Social Organization of the Sciences*. Oxford: Clarendon Press.

Wu, Stephen. 2007. "Recent Publishing Trends at the AER, JPE and QJE." *Applied Economics Letters* 14(1): 59–63.

Young, Christobal. 2009. "The Emergence of Sociology from Political Economy in the United States: 1890 to 1940." *Journal of the History of the Behavioral Sciences* 45(2): 91–116.

Zingales, Luigi. 2013. "Preventing Economists' Capture." Chap. 6 in *Preventing Regulatory Capture: Special Interest Influence and How to Limit It,* edited by Daniel Carpenter and David A. Moss. Cambridge University Press.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 115–138

The Case for Paying College Athletes

Allen R. Sanderson and John J. Siegfried

The 21st century has been both the best and worst of times for the National Collegiate Athletic Association (NCAA) and its members. Television ratings and media dollars have never been higher, owing largely to the popularity of two major revenue sports (football and men's basketball). Several college coaches have gained celebrity status and corresponding compensation packages. Among head football coaches, Alabama's Nick Saban earns over $7 million per year and Ohio State's Urban Meyer has a base salary of $4.6 million per year, not counting numerous incentive clauses. Among head basketball coaches, Duke's Mike Krzyzewski earns $9.7 million per year and Kentucky's John Calipari is above $7 million per year. The median head football coach among the 126 Football Bowl Subdivision institutions earned $1.9 million in 2013; the comparable head basketball coach's salary was $1.2 million (Fulks 2014, table 3.12(a), p. 38).

On the other hand, the NCAA has never been more vulnerable and on the defensive with regard to its policies and practices, especially its reliance on the age-old characterization of college athletes as "amateurs" who are first and foremost "student-athletes" and the limits its members have collectively imposed on the remuneration these players receive. Several recent recipients of the Heisman Trophy (an annual award given to the most outstanding college football player) have been tainted when the players were found to have received benefits

■ *Allen R. Sanderson is a Senior Lecturer in the Department of Economics, University of Chicago, Chicago, Illinois. John J. Siegfried is a Visiting Research Fellow, School of Economics, University of Adelaide, Adelaide, Australia; Secretary-Treasurer Emeritus, American Economic Association, Nashville, Tennessee; and Professor Emeritus, Vanderbilt University, Nashville, Tennessee. Their email addresses are arsx@uchicago.edu and john.siegfried@vanderbilt.edu.*

http://dx.doi.org/10.1257/jep.29.1.115 doi=10.1257/jep.29.1.115

beyond the NCAA's prescribed limits, and similar allegations are pending against both the still-active 2013 winner and a leading candidate for the 2014 Trophy. Other players have had their college eligibility revoked for violating similar NCAA rules. At colleges and universities where these players had competed, victories were vacated, football teams were banned from bowl games, and coaches were terminated. Charges of illegal payments to players, academic fraud (fake courses and plagiarism), and gross improprieties involving faculty and university administrators, arguably in attempts to protect huge athletic revenue streams, have surfaced at Penn State, Notre Dame, and the University of North Carolina-Chapel Hill, among others.

The recent explosion of revenues flowing to NCAA member institutions and the relative pittance going to the primary input—the players—for those participating in bowl games and the annual "March Madness" men's basketball tournament have created growing unease over the distribution of the largesse.[1] Much of the unseemly behavior has revolved around players seeking benefits beyond the NCAA's collectively imposed maximum compensation—formally an athletic scholarship—that is limited to tuition, room, board, books, and fees. These concerns have already led the NCAA leadership to propose modest increases in financial aid awards and to consider changing restrictions on athletes' opportunities to earn income beyond their grants-in-aid, undoubtedly an attempt to thwart demands for more far-reaching reforms that might undermine the NCAA, and completely destroy the existing intercollegiate sports business model.[2] In the next few years, several on-going legal challenges to NCAA rules will play out in ways that could alter college athletics, a uniquely American enterprise, drastically and permanently.

[1] Disclosures of lucrative financial dealings for the NCAA, leading conferences, and institutions have added fuel to the fire: a 14-year $10.8 billion contract between the NCAA, CBS Sports, and Turner Broadcasting System to televise the men's basketball tournament from 2011 to 2024; a $7.3 billion contract for the NCAA Football Bowl Subdivision playoff and six associated bowl games for the next 12 years; top-tier programs such as Texas, Alabama, Louisiana State, Oklahoma, and Nebraska, bolstered by their own or conference broadcast networks and ticket sales, showing substantial profits from athletics (Berkowitz, Upton, and Brady 2013); and the perpetual realignment and expansion of conferences that are only explainable as an attempt to capture more television revenues.

[2] More attention and scrutiny are being applied within academe itself, questioning the compatibility of big-time college athletics with the research and teaching missions of these institutions as they are reflected in admission decisions and academic practices. There are also questions about the role and power of athletic departments on campus, and the budgetary ramifications of these commitments. The NCAA has taken a beating from the media in recent years. See historian Taylor Branch's essay in the October 2011 issue of *The Atlantic* ("The Shame of College Sports"), the documentary DVD "Schooled," and harangues by Joe Nocera in the *New York Times*. In the last several years economists have also produced volumes assessing the state of athletics on college campuses, including Shulman and Bowen (2001), Bowen and Levin (2003), Clotfelter (2011), Fort and Winfree (2013), Grant, Leadley, and Zygmont (2008), and Zimbalist (1999).

A Brief History of the NCAA and its Governance

In 1906, in response to directives from President Theodore Roosevelt about the need for rule changes in intercollegiate football to reduce injury to players, several universities established the Intercollegiate Athletic Association of the United States, which became the National Collegiate Athletic Association in 1910. Concern over player health at the turn of the century was superseded 40 years later by worries about recruiting costs, the impact of television, and the financial stability of athletic programs—largely because increasing competition among colleges and universities for players was raising costs. Soon after World War II, NCAA members agreed to abide by a "Sanity Code," limiting compensation to players and setting limits on escalating recruitment costs. Walter Byers, who coined the term "student-athlete," became the NCAA's first executive director in 1951 and served until 1988 (Byers 1995); Mark Emmert, former president of the University of Washington, is the NCAA's current president. The Association, which has more than 500 employees, has been based in Indianapolis since 1999. Rent on its sizable headquarters is $1 a year, courtesy of local support for this not-for-profit organization.[3]

Subsequent forks in the NCAA's evolution have included its split into three competitive divisions—I, II, and III—in 1973, with I and II permitted to grant financial aid to athletes. In 2006, Division I further segmented football into the Football Bowl Subdivision (FBS, 126 institutions) and the Football Championship Subdivision (FCS, 122 institutions). Women's sports came under the NCAA's purview during the 1980s, precipitating decades of sparring over the Title IX amendment to the Higher Education Act of 1972—requiring gender equity in higher education.

What transformed college athletics and the NCAA from a "cottage industry" 60 years ago to the 800-pound financial behemoth it is today? First and foremost is the growth of television that fostered unprecedented expansion in broadcast revenues. Exposure via television also nudged the industry from one of local or regional interest to a national market, leading to an explosion in the number of contests and televised games, and even changes in the time of day or day of the week when they take place to accommodate endless broadcast network and cable demands for lucrative live-sports programming. While such changes have increased revenues in the affected programs, the shifting focus on college campuses toward intercollegiate sports has also had implications for academics. For example, Clotfelter (2011)

[3] In addition to developing the "student-athlete" label, the NCAA has created other now-familiar nomenclatures that apply only to intercollegiate athletes, distinguishing them from regular students. For example, television broadcasters often refer to a player as a "true freshman," as distinguished from a "freshman." In sports, the latter is actually a true sophomore who was "red-shirted" his first year—that is, he was a registered student, practiced with the team, and was in uniform during games but not allowed to play, a mechanism a coach can use to gain eligibility for the player during his fifth year in college (a "fifth-year senior"), when he is likely to be 23 years old, and much stronger and a better player than when he was 18. Unlike all other students, student-athletes essentially are classified on the basis of their sports eligibility rather than their academic progress. To our knowledge philosophy and physics departments do not red-shirt their majors, even though the students might be more accomplished scholars during their fifth year in residence. Nor do they label their first-year students as "true freshmen."

demonstrates that JSTOR usage falls substantially in March at universities whose teams are participating in the annual NCAA basketball tournaments that are played that month.

In what follows, we address the economic issue of why universities continue to operate large-scale commercial athletic programs and explore the market for players, the structure of the industry, and current legal challenges facing the NCAA. Our primary focus is on the two principal revenue sports, football and men's basketball, at high-profile athletic programs because they generate much more revenues than other sports.[4] However, changing the operations of these two programs will also have implications for nonrevenue sports, including many women's teams at these institutions. We find the prospect of a competitive labor market in big-time college athletics appealing, though how such a market would work, and the transition to it, is a challenge to envision and implement.

Why Do Universities Operate Large-Scale Commercialized Sports Programs?

Most American colleges and universities field intercollegiate athletic teams in a variety of men's and women's sports. For about 350 universities playing Division I college basketball and 126 playing football in the Football Bowl Subdivision of Division I, these contests generate substantial revenues from television broadcast rights and ticket sales.[5] Table 1 reports revenues generated by the athletic departments and subsidies from the rest of the university (called "allocated revenue" in Table 1) to cover athletic department losses for the 126 FBS universities for the last decade. The median FBS program currently operates on a budget of about $60 million per year, one-third of which is a subsidy from the rest of the university. Examples of schools with athletic department budgets near the median include Maryland, Connecticut, Mississippi State, Iowa State, Georgia Tech, and Colorado. For an institution with 20,000 undergraduates, like Georgia Tech or Mississippi

[4] In 2013, the median annual revenue generated at the 126 largest (Football Bowl Subdivision) programs from football was $20.3 million and from men's basketball $5.6 million. The next highest median revenue was $1.0 million from men's ice hockey. Beyond men's ice hockey, no other men's or women's intercollegiate sport at FBS institutions generated median revenues exceeding $600,000 in 2013.

[5] The NCAA divides its member institutions into three divisions. Division I includes roughly 350 schools that typically have the most students, the biggest athletics budgets, and the most scholarships. These schools agree to meet various minimum standards, like sponsoring at least 14 sports. Division II includes about 300 colleges and universities. In this division, there are more restricted financial aid awards for each sport, and so it is common for athletes to receive partial scholarships. Division III includes about 450 colleges and universities, often smaller in size, and while athletes at this level of competition are eligible for the same need-based financial aid as any other student, there are no athletic scholarships. Division I is further divided for football, as noted on the previous page. The Football Bowl Subdivision only includes 126 schools, which are eligible to have their football teams play in end-of-season bowl games that determine a national champion. This group of schools is subject to additional requirements: for example, these schools must average at least 15,000 in attendance at home football games. The remainder of Division I schools do not sponsor football teams.

Table 1
Financial Statistics for 126 Football Bowl Subdivision Universities, 2004–2013

	Revenue (median; millions of dollars)					
Year	*Total revenue*	*Generated revenue*	*Allocated revenue*	*Subsidy percentage*	*Total revenue 2004 dollars (average; millions of dollars)*	*# of athletes on scholarship*
2004	28.3	22.8	5.4	19.1%	28.3	577
2005	32.8	24.3	8.5	25.9%	31.6	589
2006	35.4	26.4	9.0	25.4%	32.4	588
2007	37.6	26.1	11.5	30.6%	33.5	598
2008	41.1	30.5	10.6	25.8%	34.8	602
2009	45.7	32.3	13.4	29.3%	37.9	603
2010	48.3	35.3	13.0	26.9%	39.7	611
2011	52.7	38.8	13.9	26.4%	42.3	616
2012	56.0	40.6	15.4	27.5%	44.2	615
2013	61.9	41.9	20.0	32.3%	48.2	611

Source: Fulks (2014).
Notes: All revenues are medians, reported in millions; number of athletes is average per institution. Total revenue is generated revenue plus allocated revenue, and approximately equals total expenses. Generated revenues are produced by the athletic department and include ticket sales, radio and television rights receipts, alumni contributions, guarantees, royalties, and NCAA and conference distributions. Allocated revenue comprises student fees allocated to athletics, financial transfers directly from the general fund, indirect institutional support such as payment of utilities, security salaries, etc., and direct governmental support, that is, funds from state and local government agencies designated for athletics. Subsidy percentage is allocated revenues/total revenues. Total revenue in 2004 dollars is nominal dollars deflated by the Higher Education Price Index. Data prior to 2004 are not comparable because of changes in procedures; all data since 2004 are audited.

State, the annual subsidy would be about $1,000 per student. Although median nominal revenues generated by intercollegiate sports teams have increased by 83 percent over the last decade, the growth rate in total athletic department expenses has expanded even faster, growing 115 percent over the same period (Fulks 2014, p. 12), leading to steadily growing subsidies.

The financial health of athletic departments rests on four elements: 1) the demand for television broadcast rights for live programming, 2) large, stable game attendance, 3) the desire of many universities to maintain ties with alumni and other constituents, and 4) a cartel agreement among universities to limit compensation for the essential input required to stage the games, namely the players. The first three elements boost athletic department revenues, while the fourth contains costs.

The rise in broadcast rights fees for college football and basketball games has been the main source of revenue growth of big-time college sports. Live sports target the favored audience of advertisers: 18–34 year-old males. Because few viewers record games and delete advertisements before watching and live games retain the uncertainty of outcome that recorded games lack, the demand for live sports content commands substantial broadcast rights fees, which in turn generate premium advertising rates. The relative value of this type of programming has grown as the technology for excising advertising has improved. Moreover, prime time for college football games is Saturday afternoons, when the likely alternatives are reruns of old sitcoms, low-budget infomercials, or documentaries, none of which generate much advertising revenue. In addition, the incremental cost of televising a college sports event that was going to be played regardless of broadcast status is modest, making these broadcasts especially profitable, leading to fierce competition for broadcasting rights. As a result, CBS paid about $800 million to the NCAA to televise the three-week 2014 men's basketball tournament; for comparison, in inflation-adjusted dollars, as recently as 1984 that figure was just $12 million. The new four-team football championship, together with four affiliated bowl games, commanded $610 per year from ESPN. Since 1996, the most valuable media rights for regular-season college sports events have been held by the major athletic conferences. The five dominant conferences—Atlantic Coast, Big 12, Big Ten, Pac-12, and Southeastern—control most of the attractive college football game inventory, and the regional parochialism of college football fans has left those largely geographically segregated conferences with considerable market power in the sale of their broadcast rights (Siegfried and Burba 2004).

Most of the FBS programs enjoy large gate attendance, with community, student, and alumni bases that support them almost regardless of their success on the field. Many of these teams represent land-grant universities located where there are few local competing entertainment options to undercut their pricing power. Because most college football games are played on Saturday afternoons or evenings and the fans are willing to drive many hours to attend a half dozen home games each year, the geographic reach of their fan base is large. Support of these programs also is largely independent of macroeconomic conditions, as evidenced by the continual steady climb of gate and broadcast revenues through the recession of 2007–2009.

According to a Knight Commission (2006) survey, 78 percent of Americans believe intercollegiate athletics is profitable. NCAA data, however, indicate that only 20 of the 126 Football Bowl Subdivision universities earned an operating surplus on intercollegiate athletics in 2013 (Fulks 2014, p. 13), a typical year, and only a portion of those profits were transferred to the academic side of their universities. The University of Texas is a prominent example of a school where athletics generates a profit: in 2013, Texas earned about $20 million on sports revenues of $163 million (Kirk 2014). But on average, funds flow in the opposite direction. A 2013 report in *USA Today* found that over $1 billion of student tuition and fees was transferred annually *to* athletic departments in NCAA Division I to support intercollegiate sporting ventures (Berkowitz, Upton, and Brady 2013). For example, Rutgers University subsidized athletics to the tune of $27 million in 2010 while it froze wages across the university to save $30 million. By 2013, when Rutgers was poised to enter the Big Ten Conference, its athletics subsidy had expanded to $47 million (Sargeant and Berkowitz 2014), about $1,400 for each undergraduate student. The fraction of athletic department revenues coming from the rest of the university in 2013 was 20 percent at FBS universities, 71 percent at FCS universities, and 77 percent at Division I universities that do not play football.[6] The magnitude of resources redirected from academic to athletic purposes at all but about 20 of American Division I colleges and universities is nothing short of remarkable at a time of huge legislative cutbacks in taxpayer support for public colleges and universities and incessant complaints about rising tuition that students and their families are being asked to pay.

How have roughly five out of every six of the top athletic departments persuaded their universities' presidents and boards of trustees or regents to devote scarce general funding to intercollegiate sports? After all, none of these institutions' charters mentions commercial entertainment activities in their mission statement (Clotfelter 2011). When they incur financial losses on athletics, universities seem to double down, spending ever greater amounts on salaries for coaches and improving physical facilities rather than interpreting losses as a signal to redeploy assets elsewhere. Drawing on Getz and Siegfried (2012), we identify a half-dozen possible rationales for this behavior.

First, participation in and success at intercollegiate athletics might attract larger appropriations from state legislators concerned about their constituents' perceptions of the public universities in their states, especially considering the fact that the median voter in virtually every state is not a college graduate and might be more interested in the flagship state university's football team than its library. In support of this hypothesis, Humphreys (2006) found that those institutions fielding Division I football teams among a sample of 570 public universities receive about 8 percent more taxpayer funding than otherwise comparable universities without Division I football; participation seems to matter more than success on the field. In a follow-up study, Alexander and Kern (2010) found that basketball has a similar effect for Division I programs.

Second, university athletics may increase private donations. More than a dozen studies have investigated the effects of commercialized intercollegiate athletics on private contributions to colleges and universities: some find no effect, while others report a modest positive effect (Getz and Siegfried 2012). Participation in football bowl games appears to stimulate the most contributions. Because most of the incremental donations are directed to the athletic department (Anderson 2012) and consumed by athletic department expenditures, however, it is not clear that this effect produces much benefit to the university in general.

Third, the presence of high-profile sports programs, like various other campus amenities, may attract additional applicants and enrollment. A few well-known anecdotes suggest a link from winning to applications. North Carolina State University enjoyed a 40 percent rise in applications after winning the NCAA men's basketball championship in 1983 under charismatic coach Jim Valvano. Boston College enjoyed a similar surge in applications in the aftermath of quarterback Doug Flutie's famous "Hail Mary" pass completion to win a nationally televised regular season

[6] These fractions, and all other NCAA data reported here are limited to intercollegiate athletics programs only, excluding intramural and club sport programs.

game against the then-dominant University of Miami in 1984. A recent systematic study by Pope and Pope (2009) confirms that participating in post-season competition generates additional student interest in a university, but the gains are modest and fleeting. Empirical evidence bearing on the effect of intercollegiate athletics on undergraduate application and enrollment decisions indicates that simply having Division I sports programs matters more for student recruitment than does the success of those teams, and football seems to matter more than basketball. Additional spending on intercollegiate athletics may alter the mix of institutions to which college high school seniors apply and which one they attend, but there is no evidence that intercollegiate athletics increases overall college enrollments—beyond the important but small effect of increasing the chances of some of the athletes themselves attending college (Getz and Siegfried 2012, pp. 359–63).

Fourth, spending on sports programs has the characteristics of an arms race (Frank 2004; Hoffer, Humphreys, Lacombe, and Ruseski 2014). Those few ambitious and profitable athletic departments bid aggressively for high-profile coaches and steadily improve their physical facilities to attract recruits. Small differences in expenditures can lead to large differences in success in recruiting and, subsequently, on the field. Unprofitable programs have little choice but to ratchet up their spending, or they may fall even farther behind in the competition for quality players, with potentially devastating effects on their sports revenues. In this way, the net profits of the few profitable teams steadily drive up nonplayer costs for all competitive teams, requiring universities with already unprofitable intercollegiate athletics programs to increase their subsidies.

It is important to remember that even if participation in or winning at Division I intercollegiate sports affects appropriations from state legislatures, stimulates private donations, or boosts student applications to some extent, the positive connection is not sufficient by itself to conclude that subsidizing intercollegiate athletics is wise. For example, if the purpose of investing in intercollegiate athletics is to increase contributions, one would need to demonstrate that spending an incremental $1 million on the salary for a coach stimulates more than $1 million in donations and, additionally, that it stimulates more donations than spending that same $1 million on expanded fund-raising for other worthwhile endeavors by the university development office. The same argument applies to intercollegiate athletics as a means to attract state appropriations and student applications.

Fifth, many colleges and universities set tuition well below the level sufficient to cover annual operating expenses. They selectively admit students with specific talents and characteristics (including children of financially successful alumni), and hope that some of them grow into appreciative multimillionaires willing to share their good fortune with their alma maters (Hoxby 2014). To enhance the prospects that the more successful graduates remember them during estate planning, these institutions invest in creating and maintaining emotional ties. They organize alumni cruises, send faculty to give talks to local alumni clubs, and sponsor annual “Homecoming” events that feature a football game. The challenge to the presidents of the universities is to weigh on the margin the value of funds devoted to directly and immediately improving teaching and research against the prospective value of a more visible and more successful intercollegiate athletics program that might eventually attract a sufficiently large donation to the nonathletic side of the institution in the future, such that when discounted, generates an even greater boost to teaching and research. Of course, a number of prominent institutions successfully pursue ties with their alumni without big-time commercialized sports: for example, California Institute of Technology, Carnegie-Mellon, Case Western Reserve, Chicago, Emory, Johns Hopkins, MIT, New York University, Rochester, and Washington University (St. Louis), as well the entire Ivy League.

Finally, football in particular may affect academic status in higher education. Through its role in grouping institutions into conferences, football might influence how universities view themselves and each other and how the general public perceives the place of particular institutions within higher education (Lifschitz, Sauder, and Stevens 2014). An institution's sports rivals publicly specify its peers as worthy adversaries, perhaps academically as well as athletically; peer academic assessment scores vary considerably more among than within football conferences (Lifschitz, Sauder, and Stevens 2014).

Unless one or more of the above arguments is persuasive, if universities are steadily losing money on their intercollegiate athletic programs, one might ask why they don't abandon them.[7] While the answer is not obvious, in the last century or so, only two institutions that fielded big-time football teams decided to scale back or drop them. Both the University of Chicago and Washington University (St. Louis) made that choice about the time of World War II (Clotfelter 2011) and have done quite well since abandoning Division I intercollegiate sports.

The Market for College Athletes

College sports labor markets are difficult to analyze because athlete services are heterogeneous. To deal with player skill heterogeneity, we introduce a benchmark concept of a "player skill unit." The skills of other players can be measured relative to this norm. When recruiting athletes, colleges seek to field teams with individuals who have many player skill units. The demand for player skill units is derived, depending on the incremental entertainment value created by additional player skill and the amount consumers are willing and able to pay for it. Demand is negatively sloped because of the combined effects of diminishing marginal returns to additional skill units and the usual downward-sloping demand (and corresponding marginal revenue) for entertainment.

The supply of athlete services reflects the payment necessary to induce additional player skill units into the market. It is positively sloped because additional

[7] Moreover, adding gravity to the question, some recruited athletes are admitted with weaker academic credentials than the marginal nonathlete's admissions portfolio (Shulman and Bowen 2001), risking the consequences of negative "peer effects" on the broader student body.

skill units have successively increasing opportunity costs. More skill units can be extracted from existing players through extra training. Players ask for increasingly larger compensation per additional skill unit provided because of diminishing marginal returns to training and because the marginal value of leisure time increases as it becomes scarcer. At higher compensation rates, additional skill units also can be secured from players newly attracted to the market.

In a free market for labor, universities would compete against each other for the services of new high school graduate athletes. With many universities and many high school graduates, such a market could be workably competitive. The result would be a competitive wage paid for player skills and probably a much reduced surplus earned by college athletic departments (where it is typically distributed as economic rents to department officials and to construct world-class facilities). But the NCAA and its members collectively fix college athletes' wages. Student-athletes appear to be the only category on a campus where an outside organization (the NCAA) is granted power to dictate compensation and hours of work. The American Library Association, for example, does not dictate pay levels for "student-library-workers." Moreover, financial aid packages at many doctoral programs exceed tuition and fees, including a stipend for living expenses, and graduate student stipends are not coordinated among the universities with PhD programs by an association of graduate schools.

Moreover, university athletic departments can essentially dictate many aspects of a "student-athlete's" routine, something that would not be possible if they had to obey general labor laws, such as restrictions on hours of work. Because Division I athletes have historically been considered "students" rather than employees, they are not covered by labor laws, are not eligible for workers compensation, and cannot bargain collectively via union representation.[8]

Colleges and universities deal with the prospect of hiring players in a competitive market by engineering monopsony power as a group, and then collectively agreeing to a ceiling on remuneration. It is not at all clear under what authority the NCAA specifies the number and size of athletic grants-in-aid awarded to college

[8] It is particularly surprising that in an academic environment where free and open discussion is generally encouraged that college athletes' freedom of speech is restricted more than that of professional athletes. When five St. Louis Rams players held up their hands on November 30, 2014, in a "hands up, don't shoot" reference to the August 2014 Ferguson, Missouri, death of Michael Brown and a subsequent grand-jury decision not to indict the police officer, and when prominent NBA players sported "I Can't Breathe" warm-up jerseys a few days later in reference to the death of Eric Garner on Staten Island in July 2014 and, again, the failure of a grand-jury to indict the police officer involved, no disciplinary action by their respective professional teams or leagues stifled their freedom of expression. However, a 2010 NCAA regulation forbids college athletes from expressing words, numbers, or symbols on their body or on tape attached to their body. So, writing a reference to a Biblical passage in players' eye-black (such as "Psalms 23:1") is prohibited. And to top off the irony, players often are even required to display a Nike swoosh or some other trademark when their university or coach has a contract with a trademark's owner.

football and basketball players.[9] There is no legislation, court ruling, or collective bargaining agreement that permits this coordination.

Because playing major college sports is attractive to many young men, and often is accompanied by perquisites like being a center of attention, possible future job offers from alumni, and, for a few of them, the chance of cashing in on a professional contract, there is a sufficiently elastic supply of players at a relatively low wage to fill all of the roster slots available on major college football and men's basketball teams. Division I football allows each FBS team to offer 85 football scholarships and each FCS team to offer 63, for a national total of about 19,000 football scholarships. Division I basketball includes about 350 teams at 13 scholarships each, yielding about 4,500 men's basketball grants-in-aid.

To have a low and elastic supply curve to profit from, college and university sports teams need to limit the alternatives available to the more-talented prospective players. The National Football League (NFL) and National Basketball Association (NBA) aid and abet in this regard by restricting new player entry into their leagues, limiting access to the NFL only to players three years after high school graduation and entry into the NBA only to players who have reached age 19 (a limit that soon may be raised to 20). The pool of prospective players therefore has limited alternative ways to practice, improve, and audition for the professional leagues other than to attend college. The NFL and NBA have an interest in how the NCAA operates, because universities provide free specific training, increased maturity, and reduced risk for future professional players. Moreover, because the professional leagues' collective bargaining agreements with their respective players' associations grant free-agency to players after they have been in the league for a specific number of years, delaying entry of players to a time nearer their peak playing skill saves team owners the difference between the high free-agency salaries of star players and the constrained (by the collective bargaining agreement) salaries of entry-level players. Conversely, the relationship furnishes universities with prime athletic talent at far less than competitive wage rates.

Agreements to restrict the alternatives available to prospective college athletes are essential to the NCAA's monopsony power in the athlete labor market. No organization other than the NBA and NFL specifies a minimum working age above 18 (except in a few cases where government imposes a minimum age, such as for a bartender or chauffeur). The implicit cooperation of professional sports leagues with the NCAA and its member institutions to enforce these requirements is unique.

Whether the athlete labor market reaches equilibrium at a number of players or a level of player skill units that is less than that level where supply intersects demand in a free competitive market cannot be determined. In a free market where the NCAA could not restrict roster sizes or the number of teams, the demand for

[9] In an effort to control costs, over recent decades the NCAA has progressively reduced the number of grants-in-aid that big-time football teams can offer. In September 2014, former Colorado State football kicker Durrell Chamorro sued the NCAA, challenging the current limit of 85 scholarships an FBS football team can offer as a collusive limitation that restrains trade.

the most skilled athletic labor would be higher. But in a competitive market the alternatives of prospective players would not be restricted, and so the supply curve also would require a higher wage at each level of skill unit offered. The first consideration leads to more players and skill units employed, while the second leads to fewer players and skill units employed than would otherwise occur. What is certain is that the compensation level of the college players is presently lower than it would be in a competitive market.

The pay ceiling on intercollegiate athletes leads universities to "overdose" on complementary inputs. The same institutions that have agreed not to compete on direct compensation to players instead compete furiously on the basis of other factors of production: program reputation; coach; quality of stadiums, arenas, weight-rooms, residence halls, and training-table food; scheduling games in attractive locations; and lavishing personal attention on recruits. The result is an 800-page book of NCAA rules and regulations for limiting recruiting expenses and player compensation, accompanied by a seemingly perpetual stream of scandals created by attempts to circumvent the cartel rules.

There is also an incentive to overuse underpaid inputs. When John Wooden coached UCLA basketball to ten national championships in the 1960s and 1970s, college basketball squads averaged about 25 regular-season games. The pre-tournament schedule now is 30–35 games for most teams. The college basketball season for elite programs essentially runs from October through March, the bulk of the academic year. In 1950, the regular college football season was eight games; now it is 12, with most conferences holding a championship game after the regular season.[10] As recently as 2001, there were 25 football bowl games; in 2014–15 there are 39. Thus, 62 percent of the FBS teams will play a bowl game. In addition, college football started a four-team playoff in January 2015 without reducing the number of regular-season games, which adds yet another game to the supply commitments for players on the two most successful tournament teams. There are already calls to expand the football playoffs to eight or even 16 teams, with each new round of playoffs adding yet another game to the schedules of successful teams. The 2015 NCAA national champion football team will most likely have played 15 games. Television exposure has also led to an increased number of games played at neutral sites, where both teams must travel, and to games played on weeknights during the academic year.

A chief reason for schedule expansion at the college level is that the marginal cost of the primary input in the production process is close to zero, and the players have no voice in the decision to expand the schedule, and no claim on the incremental revenues generated. In contrast, decisions to increase the number of games played by professional teams are made in consultation and agreement with the

[10] The expansion in the number of teams in college football conferences from about eight or ten to twelve or fourteen, in addition to capturing more television revenue, also facilitates adding a conference championship game, pitting the winner of one division against the other, thus sneaking in one more revenue-generating game.

players' association. As a result, the NFL has played a 16-game schedule since 1978, and the players' union blocked recent attempts to lengthen it to 18 games. The regular season in the NBA has been fixed at 82 games since 1967–68.

At the professional level, there are also safeguards regarding how long a coach can work his players, constraints imposed via negotiation between the players' association and the league. At the collegiate level there are no comparable controls over excessive hours. Although the NCAA unilaterally limits practice to 20 hours per week, there are innumerable ways coaches can circumvent the nominal limit. For example, compliance meetings, traveling to and from competitions, drug educational meetings, and community service projects do not count toward the 20-hour per week limit. Voluntary athletic-related activity in which a student-athlete participates and which is not required or supervised by coaches is also not counted against the totals. This could include strength and conditioning as well as athletic skill work. Many college football teams report for work near the end of July, one or sometimes even two months before other students return to campus from summer break.

Yet another way the NCAA stifles competition for players is by limiting their opportunity to transfer. A regular degree-seeking student who is dissatisfied with the academic or social characteristics of a particular college can transfer easily. The student's initial college cannot stop such students from leaving, nor dictate where they enroll. But the NCAA and the student-athlete's initial coach *can* dictate where a scholarship athlete may not enroll (for example, at a conference rival); plus, the player must sit out from playing for a year. No similar cost is borne by other students or coaches. A football or basketball coach who changes jobs may be required to "buy" his way out, but only if he voluntarily signed a contract containing such a stipulation. And he can begin immediately elsewhere, even before the current season is over, or before the team plays in a bowl game.

The longer one considers the NCAA-coordinated limits on what college athletes in the money-making sports can be paid and what they can do, the more uncomfortable comparisons arise. The NCAA used to fix the salaries of some assistant coaches, but a 1998 Court of Appeals ruling held that this limit was collusion in restraint of trade, an antitrust violation costing the NCAA a judgment of $66 million (*Law v. National Collegiate Athletic Association*, 134 F.3d 1010 [10th Cir. 1998]). And as noted earlier, the median head coaches of big-time football and basketball programs are paid well over $1 million per year, not the adult equivalent of "room, board, tuition, books, and fees."

The real issue is not whether college athletes should be paid, or whether all schools pay the same amount. College athletes at the Division I level are in fact currently paid, in the sense that the majority receive grants-in-aid that cover most—although not all—of their college expenses. Athletes are also paid different amounts depending on the school they attend. The NCAA policy to compensate student athletes with room, board, tuition, books, and fees masks an enormous disparity across member institutions in the dollar value of that financial aid package. For example, at Brigham Young University the full-year tuition is less than $5,000; Stanford's tuition is roughly ten times as much. One might also argue that a diploma from, or even attendance at, some colleges compared to others is worth a significant difference in terms of expected lifetime incomes.

The NCAA's Monopoly and Monopsony Power

Sixty years ago, one might not have predicted the persistent and steadily increasing market power of the NCAA. One would have expected a group of more than 1,000 institutions to have difficulty maintaining cartel stability. Moreover, NCAA members are the epitome of heterogeneity. Some are public, others are private; they vary enormously in terms of budgets, wealth, and the size and academic quality of their student bodies; and they differ by mission and their scope of activities—for example, between colleges with a predominantly teaching focus and research-oriented universities. However, despite periodic squabbles among members about how to distribute the spoils, the NCAA has been remarkably adept at creating and marketing its brand, retaining loyalties, beating back challenges to its market power, and resisting incentives for individual teams to cheat on agreements. Other than losing the 1998 assistant coaches' wage-fixing case and a 1984 US Supreme Court decision ending the collective sale of television broadcast rights (*National Collegiate Athletic Association v. Board of Regents of the University of Oklahoma* 468 US 85 [1984]), until recently the NCAA generally has prevailed in legal disputes. This legal winning streak is now in serious jeopardy, as we discuss below.

The NCAA benefits from various arrangements that allow it to exercise market power on the supply side of the market for college athletics. The range of conditions that must be met for entry means that the number of teams in the FBS and FCS of the NCAA is limited to about 250 and the number of teams in Division I for basketball to about 350. Because setting up new college sports conferences is difficult, an erosion of economic rents due to entry is of little concern to the elite. The NFL does not broadcast on Saturdays during the college football season as a result of a compromise it reached with the US Congress in the Sports Broadcasting Act of 1961, cementing college football's market power in broadcasting live sporting contests on Saturdays.

Nevertheless, the mighty edifice of big-time college athletics must still compete in selling its product with a range of other options for the consumer's discretionary time and entertainment dollar, including professional sports and nonsports options. Thus, it may be that the most important aspect of the NCAA's market power is its monopsony control over players.

The Distributional Aspects of Change: Cui Bono?

In the contemporary world of intercollegiate athletics, some parties benefit from current arrangements and others are harmed. One fact seems inescapable: rents are expropriated from the most talented football and men's basketball players in high-profile programs and redistributed to other parties. If a competitive labor market for athletes would return these rents to the players, it is important to understand who is benefiting now, because that will identify the most likely resistance to any movement toward a competitive labor market for college athletes.

One set of redistributions might be among the athletes themselves. Not all Division I football or men's basketball players currently are exploited. The star quarterback, running back, or wide receiver, or the high-scoring shooting guard or 7-foot shot-blocking center would clearly be paid more in a competitive market for college athletics talent. But a bench warmer might be paid less. The 85th grant-in-aid player on the 2014 BCS champion Florida State University football team and the last substitute on the 2014 NCAA national champion University of Connecticut basketball team bench are both likely net beneficiaries of current arrangements. The relevant question is where along the talent continuum the needle moves from exploited to subsidized.

Using conventional methodology, Lane, Nagel, and Netz (2014) measure the marginal revenue product of Division I men's college basketball players. Successively relating player performance to winning, and winning to gate receipts, they find that the playing contributions of about 60 percent of the players generate revenues exceeding the value of their grants-in-aid. For example, on most basketball teams the starting five and the first two substitutes generate net revenues, which is plausible. Those are the players likely to receive additional compensation if intercollegiate teams hired labor in a competitive market. While there is no analogous study of college football players, it is likely that 40 to 50 of the 85 scholarship players on most Division I football teams would also receive more than just a grant-in-aid in a competitive labor market. The rest would likely be worse off, particularly if more players on top Division I teams are "walk-ons," essentially nonscholarship players.

Other Division I college sports—such as wrestling, swimming, softball, and volleyball—that at most institutions do not bring in sufficient revenue from television, gate receipts, and private donations to cover their scholarships would probably be little affected by men's basketball and football players being paid a competitive market wage. Many nonrevenue sports teams at Division I universities have far more athletes, male and female, than they have full grants-in-aid, so they are in essence already treating some of the athletes in these sports like regular students, eligible only for need-based scholarships. As Fort and Winfree (2013, Chap. 1) point out, most big-time sports programs lose money, and the nonrevenue sports are already being subsidized by general university funds. However, a competitive market for football and men's basketball players could have implications for women athletes, depending on how the Title IX rules that require equity between male and female athletic scholarships are interpreted. If football players are considered employees, as the Illinois regional director of the National Labor Relations Board (NLRB) ruled in April 2014, does that remove 85 scholarships from the male side of the Title IX scales, allowing institutions to reduce female scholarships by a corresponding 85?

The effect of having the highly-recruited quarterback earning, say, $200,000 a year, with the right tackle receiving the economic value of a traditional grant-in-aid, and perhaps the English graduate assistant who is teaching both of them being paid even less does not give us pause. There already are enormous salary disparities among and within universities—as illustrated by differences in what physics and philosophy professors are paid, and the persistent arguments over the unusually low pay of adjunct faculty. Competitive markets pay workers based on their marginal revenue products and opportunity costs, and when those factors differ among individuals, compensation varies accordingly.

Another set of redistributions would presumably arise among the Division I colleges and universities with high-profile football and basketball programs. The effects could extend to shifts in intra-university transfers; shifts in authority, control, and power on their campuses; changes in the size and distribution of their applicant pools; and political costs of lobbying state legislators. Paying the players market-based wages might increase short-term financial operating losses at some—or many—universities. Those institutions with a high level of commitment to athletic excellence and a willingness to spend whatever it takes to beat their archrivals will presumably bid up the price of players. But over time, even elite programs would have to recalibrate how much they are willing to devote to paying their star performers in football or men's basketball. Such institutions would also need to consider where those monies come from—whether from academic programs, reductions in scholarships to other athletes, more fees imposed on students, larger contributions from legislatures or alumni, less spending on facilities or amenities for players, or from the salaries of the coaches and director of athletics. Otherwise, the zero-sum competitive recruiting game will drive even the highest revenue programs into bankruptcy.

We think the primary reason for the plethora of big-time university sports teams is the binding ceiling on wages paid to players. With such a distortion in factor prices, an inefficiently large number of teams can survive. It is likely that paying players would move the market for college athletics to an equilibrium of fewer teams, probably closer to the number of teams that would exist in the corresponding premier professional leagues if those leagues did not restrict entry so as to increase the value of their franchises. If the current number of high-level basketball programs were to drop from around 350 to about 100, or in football a reduction to approximately 65 programs instead of the current 126 in FBS competition (65 is the number of teams in the five "power" conferences, plus Notre Dame), then either some of those who would have been scholarship football and men's basketball players would become unemployed or work as volunteers—that is, as "walk-ons."[11] If the NFL and NBA reacted to a smaller number of big-time college athletics programs by instituting viable training-leagues, some of the potential unemployment would be mitigated. But given that the NFL and NBA mostly draft players from elite programs, and those players are most likely to survive, the professional leagues might be comfortable with a shrunken version of the college status quo, seeing little

[11] The number of FBS football teams declined by one when the University of Alabama-Birmingham announced on December 1, 2014, that it was dropping football from its athletics program because of its high cost. This is the first football program to leave the FBS in over two decades.

need to pay for training "laid off" college athletes who were unlikely to make it in the premier professional leagues anyway.

One possible outcome of paying players is that the major college conferences would break off from the NCAA entirely and conduct their athletics business in an entirely different way, including increasing the pay of players in revenue sports. In football, one could envision a world in which the five major conferences as a group, or as individual conferences, and maybe a few of the other strong conferences would reorganize into smaller cartels, and become the effective organizing unit. These cartels might pass muster with antitrust regulators, who have not challenged the conference-level coordinated sale of college television broadcast rights that developed after the Supreme Court nullified the NCAA's national broadcast cartel in 1984.

If college athletes were paid competitive market wages, how would the demand for in-venue and live broadcast game content among students, alumni, and other fans fare? Competitive balance is sometimes seen as a fundamental and necessary ingredient in any athletic contest. In his seminal sports economics article, Rottenberg (1956) wrote: "The nature of the [sports] industry is such that competitors must be of approximately equal size if any are to be successful." If college athletics moved from the current status quo to a situation that allows uncapped compensation, perhaps formally treating athletes as employees in some institutions, and reducing coordination across universities, competitive balance may change. However, it is not obvious in which direction. The existing system of capped compensation for players bestows enormous recruiting benefits on prestige programs. Institutions like Western Kentucky and the University of Massachusetts currently face an uphill battle recruiting against Notre Dame or Duke, with their high-profile programs and coaches.

How competitive balance would change if players were compensated differently would depend on the relative preferences of players for cash compensation versus their perceived value of noncash benefits of playing for various colleges or universities. Since there must be at least some highly talented players whose preferences favor cash, the introduction of pay-for-play is likely to divert some players to universities that had no chance to attract them when the recruiting currency was limited to program prestige and playing facilities.

Even if competitive balance were to decline, demand may not follow. Intercollegiate athletics currently is quite popular in spite of a fairly high degree of competitive imbalance. The demand for dominant teams and the enjoyment fans of nondominant teams receive when their team occasionally upsets a dominant team may outweigh the demand for more competitive balance (Coates, Humphreys, and Zhou 2014). After all, a few dominant teams create an opportunity for other teams to be dragon slayers. As a recent *Sports Illustrated* article put it, "without Goliath, David was just a dude throwing stones without a concealed weapons permit" (Gorant and Keith 2014).

Sports fans currently enjoy a panoply of television viewing opportunities as well as an array of in-venue intercollegiate sports options. Paying athletes would affect fans' amenities, particularly at the institutions that may reduce support for high-profile commercial athletics. When thinking about potential losses to students and alumni who are sports fans, however, it is also worth remembering that many students and alumni have little or no interest in big-time sports contests. Some even dislike sports. Such students might well prefer that they are not assessed fees to pay for such contests or that sports subsidies coming from their tuition dollars be reallocated to different extracurricular or academic activities.

There is also the fundamental question as to how paying players more, and correspondingly admitting publicly that these high-profile sports teams are comprised of hired-guns with at best only a loose affiliation to the university, might affect demand by spectators. If paying players overtly reduces the demand for viewing college sports, perhaps to levels experienced by minor leagues in baseball and ice hockey, the revenue-maximizing price fans or broadcast networks pay to watch in-person or to broadcast games on television will decline. But a simple increase in the cost of labor without any shift in demand should not affect ticket prices (Fort and Winfree 2013, chap. 10).

How athletes in nonrevenue intercollegiate sports programs would be affected hinges on how universities would rebudget if the net revenues from their football and men's basketball programs fell, forcing resources from one part of the academic or athletic enterprise to another. In most cases, however, nonrevenue intercollegiate sports are already subsidized by general university funds. These intercollegiate sports teams, as well as intramural and club sports, are part of a set of amenities institutions provide to recruit talented students and to keep them satisfied. These activities are likely to survive any sea change—except on one score: What would be the implications for Title IX and female athletes if current restrictions on football and men's basketball player compensation were eliminated? For the most part, excluding a few select high-profile women's basketball programs (like Connecticut and Tennessee), female athletes play on a wide range of low- or nonrevenue teams. On the one hand, just as with nonrevenue sports teams for men, the impact might be minimal. However, when it comes to gender equity, the interests of the federal government and the courts, as well as the institutions themselves, could turn this into a larger issue.

Next, among the many tentacles of the college sports octopus are the television and cable networks and their broadcast affiliates (an integral part of the college revenue machine); complementary firms such as Nike, Reebok, Under Armour, and other advertisers and sponsors; cities that play host to bowl games and regional March Madness weekends, whose mayors believe the events boost their local economies; and sports writers and broadcasters. They all benefit from the current overproduction of, and emphasis on, high-profile college athletics, which affords them an array of programming alternatives, inexpensive advertising, and livelihoods that depend in large part on the status quo. They are likely to be worse off in a world of pay-for-play college athletics.

When thinking about who benefits from the current arrangements, it is worth remembering that the vast majority of star Division I football and men's basketball players are African-Americans, many from low-income families. Athletes in nonrevenue sports, athletic department personnel, coaches, faculty and staff, and the student and alumni bodies of the Division I universities as a whole are predominantly white. Given that the NCAA and its members now suppress the wages of outstanding athletes to amass rents and then redistribute that largesse to other people and units on campus (as well as to the NCAA itself), the distributional implications are embarrassingly clear: lower-income (on average) minority athletes are "taxed" to provide benefits to other people who are overwhelmingly white and from higher socioeconomic strata.

One can also raise concerns that a competitive free market in college football and basketball might in some ways offer too little protection for these young men, who will find themselves (and their families) in fine-print negotiations. One can imagine a limited role for the NCAA to ameliorate these asymmetrical information problems.

How Will Change Arrive? Internal Reforms and Lawsuits

In what appears to be an effort to head off even more drastic changes in the existing intercollegiate sports business model, in April 2014, the NCAA Division I board of directors voted to allow all universities to offer unlimited meals and snacks to their athletes in addition to the restricted regular meal plans provided through their grants-in-aid, spawning a new intercollegiate competition in food provision. Subsequently, the NCAA changed its organizational structure to allow the five premier college athletic conferences and Notre Dame to operate under a different set of rules. Presumably this will allow those teams to provide additional benefits to their scholarship athletes; those benefits could include raising compensation up to the full cost of attendance at each institution, and insurance policies covering playing related medical expenses incurred after the end of a player's college career.

Other proposals now being discussed have included a requirement that schools shift to multiyear scholarships. At present, most athletic grants-in-aid are not automatically renewable from year to year, although since 2012 individual institutions may, as the University of South Carolina and the University of Southern California have done, act unilaterally to offer multiyear grants-in-aid to scholarship athletes, which essentially is a form of wage competition. As an illustration of how competition breaks out on many fronts, in October 2014, the Big Ten conference announced that henceforth all of its athletic scholarships will be guaranteed for four years. Further changes might provide support for former athletes who want to complete their undergraduate degrees after their playing eligibility has expired, or who return to school for an advanced degree. Dealing with health-related concerns that surface long after an athlete's playing days are over, such as concussions,[12] would be another possibility.

[12] In July 2014, the NCAA offered $70 million to settle claims in several head-injury-related lawsuits that are pending in US District Court in Chicago. This follows a similar dispute lodged by former professional football players against the NFL for retired players' medical costs from dementia and other neurological disorders tied to repeated concussions.

Whatever the merits of these proposals, they fall well short of a free competitive labor market for college athletes. It seems likely that any change beyond tinkering will require the pressure of government regulatory decisions and lawsuits. Several pending lawsuits seem especially salient.

First, there is the 2009 complaint in *O'Bannon v. NCAA* that the trial court decided in August 2014. Ed O'Bannon, a former player on UCLA's last national championship basketball team, argued that after players leave college they should share revenues from the commercial use of their image; the NCAA has asserted lifetime control over those rights. US District Court Judge Claudia Wilken ruled in the O'Bannon case that the NCAA's collective agreement to cap player compensation at the level of a grant-in-aid violates the Sherman Act because it is collusion in restraint of trade. To complicate matters, however, Judge Wilken went on to suggest that a compensation cap set above the current level of tuition, room, board, books, and fees (by $5,000) might withstand legal scrutiny. The ruling (which can be downloaded here: http://s3.documentcloud.org/documents/1272774/obannon-court-decision.pdf) is under appeal.[13]

Second, several former Northwestern University athletes recently have organized the College Athletes' Players Association, which argues that college players are employees who should be eligible for employee medical benefits and allowed to bargain collectively over compensation and work conditions. In March 2014, a regional director of the National Labor Relations Board ruled that Northwestern's football players are primarily employees, rather than "student-athletes" as the NCAA maintains. Although the NCAA and Northwestern are appealing this decision, we believe that at least some Division I schools eventually might welcome a union representing college players. With a union in place, the teams in the conferences or even groups of conferences could negotiate a collective bargaining agreement similar to the agreements between professional sports leagues and their players' associations that include various provisions that would otherwise be illegal, like maximum and minimum salary levels, or a team payroll cap.

Third, a collection of similar cases that directly attack the ceiling on grants-in-aid are moving through the courts. One prominent suit filed on behalf of former running back Shawne Alston requests that he be paid the foregone earnings he might have earned from West Virginia University if the school had not agreed with other NCAA colleges and universities to restrict his compensation to a grant-in-aid. To add fuel to the fire, in March 2014, prominent sports labor attorney Jeffrey Kessler filed a class-action lawsuit (Farrey 2014) in a federal court in New Jersey against the NCAA and its five "power conferences."[14] Interestingly, Kessler's lawsuit

[13] The success of O'Bannon may have precipitated further liability for television sports broadcast networks and the college conferences with which they negotiate broadcast rights. On October 3, 2014, ten more former college athletes sued numerous broadcast networks and athletic conferences who have profited from the broadcast and use of student athletes' names, likenesses, and images without the athletes' permission (*Jevon Marshall et al. v. ESPN, Inc.*, Case 3:14-cv-01945, Middle District of Tennessee).

[14] The plaintiff's name in the Kessler case is Martin Jenkins, a former Clemson football defensive back, but the name most often mentioned is the lawyer Kessler because he is a formidable legal opponent in sports labor matters. Two decades ago, Kessler won the case that led to free agency for NFL players.

does not ask for financial damages but instead for an injunction that would eliminate all collectively imposed restrictions on player compensation. The request for an injunction, rather than damages, suggests the case is based on principle rather than a quest for financial reward and is therefore less likely to settle out of court by compromise. Such cases are straightforward—they ask the courts to find the collective NCAA restrictions on the number and size of grants-in-aid illegal under the Sherman Act, and to issue an injunction against the organization and its five largest conferences prohibiting them from continuing the practice.

It is of course impossible to forecast the eventual outcomes of these cases. But the precedents from the NCAA's legal defeats mentioned earlier—both the 1984 television broadcast rights price-fixing case and the 1998 assistant coaches wage-fixing case—suggest the NCAA is in risky legal territory with respect to its agreement to limit player compensation. The enormous increase in revenues for Division I football and basketball in the last few decades has fundamentally altered the question of whether it is reasonable for player compensation to be limited to grants-in-aid. These pending lawsuits are likely to lead to changes well beyond the incremental steps currently proposed by the NCAA.

We expect an evolution in the labor market for big-time college athletes, primarily in the form of changes that greatly reduce, if not completely eliminate, the monopsony power of the NCAA, intercollegiate sports teams, and conferences. How long will it take to reach a new steady-state equilibrium? Is it possible to reach the new equilibrium with only modest disruption to the existing structure of revenue-producing college athletics? Or is considerable confusion and ensuing chaos part and parcel of the athletic bed the NCAA has made for itself? Our sense is that people involved in big-time intercollegiate athletics are too ambitious and too aggressive to control themselves unilaterally so as to operate within the constraints of antitrust law.

The current arrangements in the labor market for big-time college athletics are inefficient, inequitable, and very likely unsustainable. Yet it is far from obvious how to get from here to a competitive labor market without incurring substantial transition costs. While a truly "competitive free market" is attractive, it is not without risk, especially considering that the output restrictions arising from big-time intercollegiate sports teams' market power in selling tickets and broadcast rights might have been offsetting the expansive pressures of the low price of labor. Second-best considerations might be important here; for example, eliminating monopsony power could lead to increased market distortions when there is no longer a force offsetting the surviving output market power.

Professional sports experienced an evolutionary process in moving toward more competitive labor markets. Professional team owners initially had total monopsony power over players. The players gradually gained somewhat equal footing through court decisions and unionization. We might expect a similar evolution in college athletics, primarily through changes that slowly erode the monopsony power of the teams and conferences. Labor discussions in the NFL or NBA now consist of the commissioner, team representatives, and a battery of lawyers on one side of the table and players, their union representatives, and their attorneys on the other side. Perfect competition this is not, but a fair fight it arguably is. In the current collegiate counterpart, on one side of the table is the athletic director, the head coach, the NCAA, and legal expertise, and on the other a 17-year-old kid and his mom; it's not hard to predict that outcome. Salary negotiations in a competitive free market for labor services would probably still involve the athletic director, the head coach, and the university's lawyers on one side of the table, but this time the 17-year-old kid and his mom are likely to be accompanied by their attorney, perhaps working on a contingency basis linked to the salary negotiated.

What might happen if Kessler ultimately prevails and college athletes can sell their athletic services in a truly free market? With 65 FBS teams and many aggressive coaches, it seems inevitable that as soon as the compensation limit is lifted, some universities will begin to offer their better players financial inducements to stay on their team and will begin to include a cash payment in packages offered to new recruits. As some institutions do so, others will follow suit. The NCAA and its members probably can tolerate an eventual O'Bannon victory upheld on appeal so long as the $5,000 supplemental per player cash payment cap outlined by Judge Wilken survives. An ultimate decision affirming the Northwestern NLRB ruling that college football players are employees would be harder, though not impossible, for the NCAA to stomach. An eventual victory by the plaintiffs in Kessler's case probably ends business-as-usual. NCAA President, Mark Emmert, when asked recently about Kessler's lawsuit said it would "blow up college sports" (Strauss 2014).

At least initially, other excessive costs that have been absorbing the rents created by the players are unlikely to diminish. So costs of big-time athletics programs will rise and the surpluses for the 20–25 programs that are currently profitable will begin to fade. The subsidies from the general fund to the athletic departments at institutions currently reporting a loss will increase. University presidents will have to confront difficult questions: "How much is too much of a subsidy? When do the benefits from fielding a competitive FBS football or a March Madness tournament-quality basketball team begin to fall short of the value of the research and teaching sacrificed to support the team financially?" It seems unlikely that the landscape of big-time commercialized intercollegiate athletics 10 years from now will resemble today's incarnation, or anything seen in the last half-century.

■ *The authors express appreciation to T. Aldrich Finegan and Lester Munson for helpful conversations, to Jack Cinoman for stellar research efforts, and to David Autor, Robert Baade, Burton Bogitch, Chang-Tai Hseih, Brad Humphreys, Richard Pomfret, Caroline Siegfried, and Timothy Taylor for helpful comments on earlier drafts.*

References

Alexander, Donald, and William Kern. 2010. "Does Athletic Success Generate Legislative Largess from Sports-Crazed Representatives? The Impact of Athletic Success on State Appropriations to Colleges and Universities." *International Journal of Sport Finance* 5(4): 253–67.

Anderson, Michael L. 2012. "The Benefits of College Athletic Success: An Application of the Propensity Score Design with Instrumental Variables." NBER Working Paper 18196.

Berkowitz, Steve, Jodi Upton, and Erik Brady. 2013. "Most NCAA Division I Athletic Departments Take Subsidies." *USA Today,* July 1.

Bowen, William G., and Sarah A. Levin. 2003. *Reclaiming the Game: College Sports and Educational Values.* Princeton University Press.

Branch, Taylor. 2011. "The Shame of College Sports." *The Atlantic.* October.

Byers, Walter. 1995. *Unsportsmanlike Conduct: Exploiting College Athletes.* University of Michigan Press.

Clotfelter, Charles T. 2011. *Big-Time Sports in American Universities.* Cambridge University Press.

Coates, Dennis, Brad R. Humphreys, and Li Zhou. 2014. "Reference-Dependent Preferences, Loss Aversion and Live Game Attendance." *Economic Inquiry* 52(3): 959–73.

Farrey, Tom. 2014. "Jeffrey Kessler Files Against NCAA." *ESPN.com.* March 18.

Fort, Rodney, and Jason Winfree. 2013. *15 Sports Myths and Why They're Wrong.* Stanford University Press.

Frank, Robert H. 2004. "Challenging the Myth: A Review of the Links Among College Athletic Success, Student Quality, and Donations." Prepared for the Knight Foundation Commission on Intercollegiate Athletics, May.

Fulks, Daniel L. 2014. *NCAA Division I Intercollegiate Athletics Programs Report, 2004– 2013: Revenues and Expenses.* The National Collegiate Athletic Association: Indianapolis, Indiana. http://www.ncaapublications.com/p-4344-division-i-revenues-and-expenses-2004-2013.aspx.

Getz, Malcolm, and John Siegfried. 2012. "What Does Intercollegiate Athletics Do To or For Colleges and Universities? Chap. 19 in *Handbook on Sport Economics,* edited by L. Kahane and S. Schmanske. Oxford University Press.

Gorant, Jim, and Ted Keith. 2014. "Big Game Hunting." *Sports Illustrated,* May 26, p. 17.

Grant, Randy R., John Leadley, and Zenon Zygmont. 2008. *The Economics of Intercollegiate Sports.* World Scientific Publishing Company.

Hoffer, Adam, Brad R. Humphreys, Donald J. Lacombe, and Jane E. Ruseski. 2014. "The NCAA Athletics Arms Race: Theory and Evidence." Working Paper, Department of Economics, West Virginia University.

Hoxby, Caroline. 2014. "The Economics of Online Postsecondary Education: MOOCs, Nonselective Education, and Highly Selective Education." *American Economic Review* 104(5): 528–33.

Humphreys, Brad R. 2006. "The Relationship between Big-time College Football and State Appropriations for Higher Education." *International Journal of Sport Finance* 1(2): 119–28.

Kirk, Jason. 2014. "College Athletic Departments Aren't Necessarily as Broke as You Think." SBNation.com, June 6.

Knight Foundation Commission on Intercollegiate Athletics. 2006. *Public Opinion Poll.* January. http://knightcommission.org/images/pdfs/pollresults1-20-06.pdf.

Lane, Erin, Juan Nagel, and Janet S. Netz. 2014. "Alternative Approaches to Measuring MRP: Are All Men's College Basketball Players Exploited?" *Journal of Sports Economics* 15(3): 237–62.

Lifschitz, Arik, Michael Sauder, and Mitchell L. Stevens. 2014. "Football as a Status System in U.S. Higher Education." *Sociology of Education* 87(3): 204–219.

Pope, Devin G., and Jaren C. Pope. 2009. "The Impact of College Sports Success on the Quantity and Quality of Student Applications." *Southern Economic Journal* 75(3): 750–80.

Rottenberg, Simon. 1956. "The Baseball Players' Labor Market." *Journal of Political Economy* 64(3): 242–58.

Sargeant, Keith, and Steve Berkowitz. 2014. "Subsidy of Rutgers Athletics Jumps 67.9% to $47 Million." *USA Today,* February 23.

Shulman, James L., and William G. Bowen. 2001. *The Game of Life: College Sports and Educational Values.* Princeton University Press.

Siegfried, John J., and Molly Gardner Burba. 2004. "The College Football Association Television Broadcast Cartel." *Antitrust Bulletin* 49(3): 799–819.

Strauss, Ben. 2014. "After Ruling in O'Bannon Case, Determining the Future of Amateur Athletics." *New York Times,* October 21.

Zimbalist, Andrew. 1999. *Unpaid Professionals: Commercialism and Conflict in Big-Time College Sports.* Princeton University Press.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 139–162

Pricing in the Market for Anticancer Drugs[†]

David H. Howard, Peter B. Bach, Ernst R. Berndt, and Rena M. Conti

In 2004, Genentech introduced the drug bevacizumab—brand name Avastin—for patients with late-stage colorectal cancer. The drug cost $50,000 per treatment episode and was associated with an incremental increase in life expectancy of five months. Following Genentech's pricing announcement, newspapers ran stories with titles like "Cancer Weapons, Out of Reach" in the *Washington Post* (Wittes 2004) and "Price of Cancer Drugs Called 'Mind-Boggling'" in *USA Today* (Szabo 2004). Some Wall Street analysts worried that bevacizumab's pricing would prompt the US Congress to regulate drug prices (Anand 2007). By 2011, the backlash against bevacizumab was a distant memory. Bristol-Myers Squibb set the price of its newly approved melanoma drug ipilimumab—brand name Yervoy—at $120,000 for a course of therapy. The drug was associated with an incremental increase in life expectancy of four months.

■ *David H. Howard is Associate Professor, Department of Health Policy and Management, Rollins School of Public Health, and Department of Economics, Emory University, Atlanta, Georgia. Peter B. Bach is a Member in the Department of Epidemiology and Biostatistics, Attending Physician in the Department of Medicine, and Director of the Center for Health Policy and Outcomes, Memorial Sloan Kettering Cancer Center, New York City, New York. Ernst R. Berndt is the Louis E. Seley Professor in Applied Economics, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts. Rena M. Conti is Assistant Professor of Health Policy, Departments of Pediatrics and Public Health Sciences, University of Chicago, Chicago, Illinois. Their email addresses are david.howard@emory.edu, bachp@mskcc.org, eberndt@mit.edu, and rconti@uchicago.edu.*

[†] To access the Appendix, Data Appendix, and disclosure statements, visit
http://dx.doi.org/10.1257/jep.29.1.139

doi=10.1257/jep.29.1.139

Drugs like bevacizumab and ipilimumab have fueled the perception that the launch prices of new anticancer drugs and other drugs in the so-called "specialty" pharmaceutical market have been increasing over time and that increases are unrelated to the magnitude of the expected health benefits (Experts in Chronic Myeloid Leukemia 2013; Kantarjian, Fojo, Mathisen, and Zwelling 2013; Schrag 2004; Hall 2013). A commentary in *The Lancet*, a leading British medical journal, summarized the conventional wisdom: "[T]he cost of the new generation of drugs is getting out of all proportion to the added benefit" (Cavalli 2013). The public debate has focused on a handful of high-profile drugs like bevacizumab. It is unclear in these debates whether these drugs are outliers or reflect broader trends in the industry.

In this paper, we discuss the unique features of the market for anticancer drugs and assess trends in the launch prices for 58 anticancer drugs approved between 1995 and 2013 in the United States. Drugs used to treat other conditions have also been closely scrutinized—most recently the $84,000 hepatitis C treatment Sovaldi—but we restrict attention to anticancer drugs because the use of median survival time as a primary outcome measure provides a common, objective scale for quantifying the incremental benefit of new products.

The market for anticancer drugs is economically significant. Within the market for pharmaceuticals, anticancer drugs rank first in terms of global spending by therapeutic class: $91 billion in 2013, up from $71 billion in 2008 (IMS 2014). The US market size was $37 billion in 2013, of which one-third was spent on 10 patent-protected cancer drugs alone (Conti, Bernstein, Villaflor, Schilsky, Rosenthal, and Bach 2013). The market is also politically salient. Anticancer drugs figure prominently in discussions over health reform, alternately symbolizing wasteful spending and biomedical progress.

We find that the average launch price of anticancer drugs, adjusted for inflation and health benefits, increased by 10 percent annually—or an average of $8,500 per year—from 1995 to 2013. We review the institutional features of the market for anticancer drugs, including generous third-party coverage that insulates patients from drug prices, the presence of strong financial incentives for physicians and hospitals to use novel products, and the lack of therapeutic substitutes. We argue that under these conditions, manufacturers are able to set the prices of new products at or slightly above the prices of existing therapies, giving rise to an upward trend in launch prices. Government-mandated price discounts for certain classes of buyers may have also contributed to launch price increases as firms sought to offset the growth in the discount segment by setting higher prices for the remainder of the market.

Drug Pricing Strategies

The process by which firms establish the "launch prices" of new, branded drugs—that is, the prices firms set immediately following US Food and Drug Administration (FDA) approval—is opaque, and relatively little work has been done on the subject.[1] At the time of FDA approval, most drugs are on-patent, and so manufacturers are temporary monopolists. They have wide leeway, though not unlimited power, to set prices.

Reekie (1978) and Lu and Comanor (1998) studied the determinants of drugs' launch prices for drugs across multiple therapeutic categories. They found that prices are higher for drugs that offer significant benefits compared to existing products. Hedonic pricing studies of colorectal cancer (Lucarelli and Nicholson 2009) and anti-ulcer drugs (Suslow 1996; Berndt, Bui, Reiley, and Urban 1995) find that manufacturers set higher prices for higher-quality drugs, but studies of antidepressants (Chen and Rizzo 2012) and arthritis drugs (Cockburn and Anis 2001) actually find the opposite. In most therapeutic categories physicians and patients learn about drug quality partly through experience, and so manufacturers may find it advantageous to introduce high-quality drugs at low prices so that the drugs will penetrate the market more quickly (Chen and Rizzo 2012).

Anticancer Drugs

Anticancer drugs are among the only life-prolonging treatments available for patients with metastatic tumors, which means that the tumor has spread beyond its original site to a nonadjacent location. The vast majority of patients with metastatic disease will die of cancer. It has become increasingly common to administer anticancer drugs to patients with early-stage disease after they have undergone surgery or radiotherapy. Because most newly approved anticancer drugs are approved on the basis of their effectiveness in patients with metastatic disease, our analysis focuses on this group of patients.

Rapid progress in the fields of tumor biology, genetics, and immunology has spurred the development of a number of new anticancer drugs. Almost 1,000 anticancer drugs are currently in various phases of pre-approval testing, more than the number for heart disease, stroke, and mental illness combined (IMS 2014; PhRMA 2014). Many new drugs are approved for the treatment of tumors with particular genetic markers. For example, the FDA approved pertuzumab in 2012 for patients with metastatic breast cancer linked to a defective HER2 gene. Targeted therapies are more likely to succeed in clinical trials and may face a less-elastic demand curve, facilitating premium pricing (Trusheim and Berndt 2012).

The scientific knowledge embodied by new drugs is impressive, but progress in basic science has not always been accompanied by proportionate improvements in patient outcomes. Gains in survival time associated with recently approved anticancer drugs are typically measured in months, not years.

[1] Prior work on pricing in the pharmaceutical industry has mostly focused on the effect of generic competition on price levels (for example, Caves, Whinston, and Hurwitz 1991) and post-entry pricing dynamics (Lu and Comanor 1998).

Most anticancer drugs are approved by the FDA on the basis of one or more randomized controlled trials. Some trials have an “active control”; patients are randomized to receive the new drug or an alternative therapy. When a drug is sufficiently novel that it has no close substitutes or it will be used in combination with existing drugs, patients in the control arm may be randomized to receive the new drug or a placebo. Trials of anticancer drugs usually measure patient outcomes in terms of the difference in survival between the treatment and control arms.

Some drugs are approved on the basis of single-arm trials. In a single-arm trial, all patients receive the new drug. There is no control group. Single-arm trials focus on short-term patient safety rather than patient survival, and so they have a much shorter duration. The FDA grants approval for many leukemia and lymphoma drugs on the basis of single-arm trials. Median survival among patients with these types of cancers is two or more years. Requiring manufacturers of leukemia and lymphoma drugs to conduct randomized trials to measure survival benefits could significantly delay the introduction of potentially beneficial drugs. Single-arm trials can show that a drug is safe but cannot determine whether the drug improves life expectancy. Physicians can observe survival in their own patient populations, but it is probably difficult for individual physicians to draw sound inferences about the quality of a new drug because their patient panels are not sufficiently large. Unlike single-arm studies, randomized trials establish efficacy as common knowledge.

Economists have measured the value of anticancer drugs by evaluating changes in life expectancy and costs over time (Howard, Kauh, and Lipscomb 2010; Lichtenberg 2009a, b; Sun, Jenna, Lakdawalla, Reyes, Philipson, and Goldman 2010; Woodward, Brown, Steward, Cronin, and Cutler 2007) or measuring patients’ willingness-to-pay (Goldman, Jena, Lakdawalla, Malin, Malkin, and Sun 2010; Lakdawalla, Romley, Sanchez, Maclean, Penrod, and Philipson 2012; Romley, Sanchez, Penrod, and Goldman 2012; Seabury, Goldman, Maclean, Penrod, and Lakdawalla 2012; Snider, Romley, Vogt, and Philipson 2012). A common finding is that the dollar-denominated benefits associated with anticancer drugs are equal to or exceed the cost of an episode of treatment. However, willingness-to-pay estimates must be interpreted cautiously in light of the fact that most patients mistakenly believe that anticancer drugs cure cancer (Weeks et al. 2012). In addition, these past studies do not address trends in launch prices. If new drugs have higher prices per unit of benefit, then we cannot assess the cost-effectiveness of anticancer drugs as a class based on studies of older drugs.

Policies Governing Drug Coverage and Reimbursement

Medicare is the most prominent US payer for anticancer drugs, followed by commercial insurers and then state Medicaid programs. Medicare pays for physician-administered intravenous drugs through the medical “Part B” benefit. By law, Medicare does not directly negotiate with drug manufacturers over prices for prescription drugs covered under the Part B benefit or the oral anticancer drugs covered under Medicare's pharmacy "Part D" benefit. Section 1861 of the Social Security Act, which requires that the Medicare program cover "reasonable and necessary" medical services, precludes consideration of cost or cost-effectiveness in coverage decisions (Neumann 2005). Consequently, Medicare covers all newly approved anticancer drugs for indications approved by the FDA.

The private insurance plans that provide prescription drug coverage under Medicare "Part D" are required to cover all drugs in six protected classes, one of which is anticancer drugs (Center for Medicare and Medicaid Services 2014). Three quarters of the population reside in states that require insurers to cover anticancer drugs for "off label," non-FDA-approved uses (Bach 2009).

Insurers in states without these requirements and large employers that self-insure have more leeway to determine coverage policies, yet, in the rare instances where third-party payers have tried to place meaningful restrictions on patients' access to anticancer drugs, they have relented under pressure from clinicians and patient advocacy groups. In the early 1990s, many insurers refused to cover a breast cancer treatment consisting of higher-than-normal doses of anticancer drugs followed by a bone marrow transplant. Breast cancer patient advocacy groups waged a high-profile campaign to secure coverage, and most insurers started paying for the treatment. Randomized trials later found that it did not prolong survival, and physicians and patients abandoned the procedure (Howard et al. 2011).

Oregon's Medicaid program recently proposed to limit coverage of anticancer drugs on the grounds that "in no instance can it be justified to spend $100,000 in public resources to increase an individual's expected survival by three months when hundreds of thousands of Oregonians are without any form of health insurance" (as reported in Landsem 2013). The proposal was withdrawn following a public backlash.

The case of bevacizumab illustrates the laxity of payers' coverage policies. The FDA approved the drug for the treatment of colorectal cancer in 2004 and then for treatment of breast cancer in 2008 based on the results of a randomized trial. Results from two additional randomized trials were later released in 2009. The trials found that patients receiving bevacizumab experienced a statistically significant gain in "progression-free survival," which measures the period of time where the cancer is under control, but that differences in overall survival were small and not statistically significant. Based on these findings, the FDA revoked coverage for bevacizumab's breast cancer indication in 2011. However, an expert panel convened by the National Comprehensive Cancer Network (2010), a consortium of major cancer centers, voted against removing bevacizumab from its list of appropriate breast cancer drugs. Faced with these conflicting decisions, Medicare and major multistate insurance plans announced they would continue to cover bevacizumab for breast cancer patients.

Some drug industry critics hold up the British National Health Service as a model for restraining drug prices. Britain's National Institute for Clinical Effectiveness evaluates the cost-effectiveness of new drugs and has restricted National Health Service funding for cancer drugs where the benefits are small in relation to costs. The British government uses the threat of noncoverage to negotiate discounts with drug manufacturers. However, restrictions on patient access are unpopular, and Prime Minister David Cameron created a 200 million pound Cancer Drugs Fund in 2011 to pay for noncovered cancer drugs outside of normal funding channels (Fleck 2013).

The oncologists who provide care to cancer patients face financial incentives to administer intravenous anticancer drugs. In most industries, there is not much difference between wholesale and retail prices, and so these prices send consistent signals. But wholesale and retail prices for drugs can diverge systematically, providing incentives for dysfunctional behavior. Oncologists and hospitals buy intravenous, physician-administered drugs from wholesalers and bill insurers. They profit on the spread between the reimbursed price and the wholesale cost. Medical oncology practices derive more than 50 percent of their revenues from drugs (Akscin, Barr, and Towle 2007), and many oncologists report that they face financial incentives to administer anticancer drugs (Malin, Weeks, Potosky, Hornbrook, and Keating 2013). Oncologists' drug choices are responsive to profit margins (Conti, Rosenthal, Polite, Bach, and Shih 2012; Jacobson, O'Malley, Earle, Pakes, Gaccione, and Newhouse 2006; Jackobson, Earle, Price, and Newhouse 2010). The use of irinotecan—brand name Camptosar—decreased following the expiration of its patent, even though the price dropped by more than 80 percent, possibly reflecting declines in the spread between the reimbursement level and oncologists' acquisition cost (Conti et al. 2012).

Insurers use cost-sharing—that is, copayments, coinsurance, and deductibles—to make patient demand responsive to the cost of health care, but cost sharing is not always effective in reducing patients' demand for anticancer drugs. Most employer-based insurance policies have an annual out-of-pocket maximum, beyond which the insurer assumes 100 percent of the cost of care. Many patients with late-stage cancer reach the maximum fairly quickly, in which case the insurer bears the full cost of anticancer drugs for the remainder of the benefit year.[2] Consequently, patients may be indifferent between a drug that costs $20,000 and one that costs $100,000.

An analysis of private insurance claims data from 1997 to 2005 found that the annual median out-of-pocket cost for the intravenous drug rituximab was $431 per year (Goldman et al. 2010). Patients' costs were less than 2 percent of total spending on rituximab. Patients' out-of-pocket costs for oral agents, which are covered under insurers' pharmacy benefit, are higher. Still, a separate analysis of claims found that cancer patients' out-of-pocket costs were 5 percent of total drug costs, and only 34 percent of patients faced per claim copayments in excess of $50 (Raborn, Pelletier, Smith, and Reyes 2012).

Even when patients face large out-of-pocket costs for anticancer drugs, they have several options for reducing their liabilities. Patients with private insurance can apply for aid from drug manufacturers' co-pay assistance programs, which offset patients' out of-pocket costs, typically on generous terms. For example, Dendreon's

[2] In the past, some plans did not count spending on prescription drugs towards the out-of-pocket maximum, but this practice is prohibited by the Patient Protection and Affordable Care Act of 2010, beginning in 2014.

patient assistance program covers up to $6,000 of patients' copayments, coinsurance, and deductibles for its $93,000 prostate therapy sipuleucel-T, boasting "75 percent of patients receiving Provenge [the trade name for sipuleucel-T] are expected to have minimal to no out-of-pocket costs" (Dendreon 2014). The program even reimburses patients for the costs they incur during travel to oncology clinics. These funds flow directly from pharmaceutical companies to patients and are not captured in insurers' records. Patient assistance programs lower the elasticity of patient demand, enabling manufacturers to set higher prices (Howard 2014). The federal government does not allow assistance programs affiliated with a pharmaceutical manufacturer to aid Medicare and Medicaid enrollees on the grounds that these programs provide an illegal inducement for patients to receive care, but manufacturers are allowed to donate funds and steer Medicare and Medicaid patients to programs operated by independent foundations. Patients can also use death as a backstop against medical debt. Most patients considering whether to use anticancer drugs have short life expectancies. They may be willing to exhaust their assets to buy small gains in health. Health care providers must write-off debt in excess of the decedent's estate.

Not surprisingly, the elasticity of demand with respect to patients' out-of-pocket costs is low. Goldman et al. (2006) estimate that spending on cancer drugs declines by 0.1 percent in response to a 10 percent increase in patient coinsurance. For the sake of comparison, spending on drugs used to treat arthritis declines by 2.1 percent and spending on drugs used to treat kidney failure declines by 0.7 percent when patient coinsurance increases by 10 percent.

Trends in Launch Prices

We evaluate pricing trends for 58 anticancer drugs approved in the US between 1995 and 2013 (CenterWatch 2014). We restrict attention to drugs administered with the primary intent of extending survival time for cancer patients and drugs for which survival benefits have been estimated in trials or modeling studies. We do not consider drugs administered to treat pain or drugs that are administered to alleviate the side effects of cancer treatments. Details about the selection of drugs, references for survival benefits, and other details about the data are provided in an Appendix available with this paper at the journal's website, http://e-jep.org.

The FDA approves drugs for specific uses, or indications, which are described in each drug's "product label." We focus on the benefits associated with each drug's first FDA-approved indication. Once a drug is FDA-approved, physicians are free to use the drug for any patient with any condition, but manufacturers may not promote the drug for "off label" indications. We did not consider the survival benefits associated with indications approved by the FDA after the initial approval of the drug. In most cases, the benefits associated with these indications are unknown to manufacturers at the time of launch and are thus difficult to incorporate into their initial pricing decisions.

Forty-one of the 58 drugs in our sample were approved on the basis of randomized controlled trials. We obtained information on the incremental survival benefits of these drugs from the results of these trials. Drugs are typically tested against the next-best therapy available at the time the trial was initiated. In some cases the next-best therapy is "nothing," and so patients receive a placebo. We measured benefits by subtracting median overall survival in the control arm from median overall survival in the treatment arm. We used progression-free survival (the period of time the cancer is under control) when trials did not report overall survival.[3] Drug manufacturers may focus on progression-free survival for practical reasons. Trials designed to detect differences in progression-free survival are shorter (progression precedes death) and require a smaller sample size because the variation in progression-free survival is typically lower than the variation in overall survival. There is considerable debate in the oncology community about whether progression-free survival is a good proxy for overall survival. Our view is that even if progression-free survival benefits are only weakly correlated with overall survival benefits, data on progression-free survival benefits provide a useful signal of product quality to a manufacturer who must set a price for a new drug in the absence of information on overall survival benefits and to practicing physicians who must decide whether to use it. In our data, we observe both overall survival and progression-free survival for 20 drugs. The absolute difference between overall survival and progression-free survival is less than one month for five of these drugs and less than two months for 13 of the drugs.

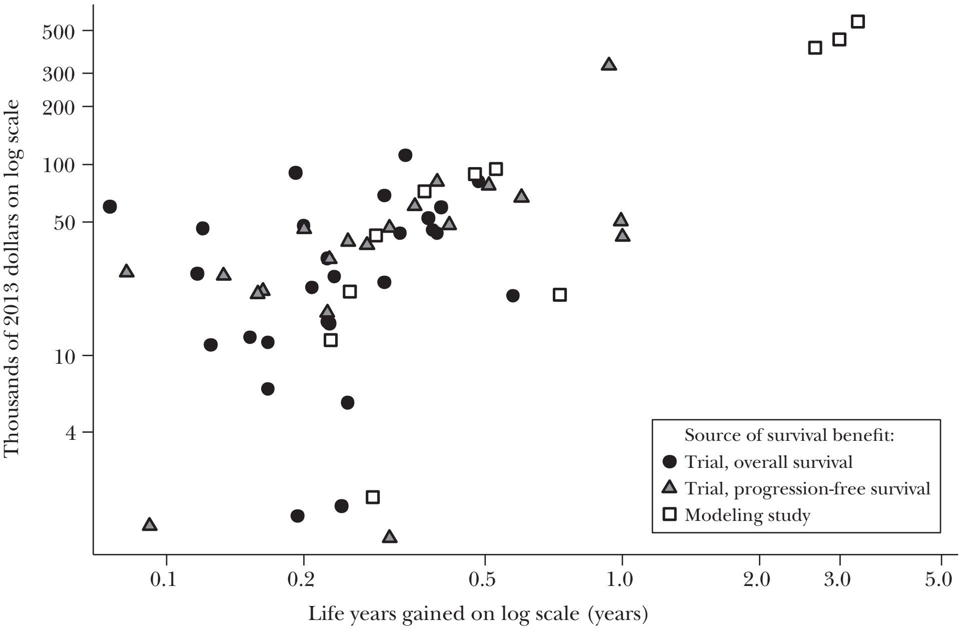
For the 17 drugs that were approved on the basis of single-arm trials, we obtained estimates of survival benefits from post-approval trials ($N = 6$) and cost-effectiveness studies that use simulation models to project survival ($N = 11$). Cost-effectiveness studies typically report benefits in terms of mean life expectancy or mean quality-adjusted life-years. We converted these quantities to median survival gains assuming survival time is distributed exponentially.[4]

We calculated the "episode treatment price" for each drug, which equals each drug's monthly cost to the Medicare program in 2013 dollars (see Bach 2009 for details) multiplied by the typical duration of treatment in months. Medicare costs represent the actual dollar amounts Medicare, the largest public insurance program, pays for drugs. In most cases, Medicare reimbursements will be greater than the prices hospitals, physicians, and pharmacies pay to wholesalers. We do not believe that rebates—refunds from manufacturers to hospitals, physicians, pharmacies, and third party insurers—are large in the market for new anticancer drugs, but pricing is opaque and rebate arrangements are closely guarded. Medicare has adjusted its payment formulae over time to align reimbursement and wholesale prices more closely. For this reason, our price series may understate increases in providers' acquisition prices. As we describe below, drug acquisition costs vary

[3] Trials report medians, because measurement of means is possible only after all patients in the trial are dead. Some trials are not powered to detect changes in overall survival but report it anyway.

[4] If we assume survival time is distributed exponentially, it is possible to convert means to medians without estimating ancillary shape parameters. Median survival is equal to mean survival multiplied by $\ln(2)$.

Figure 1
Drug Prices versus Life Years Gained



Source: Authors.

between providers and pharmacies, and Medicare payment rates do not account for differences in acquisition costs across various categories of buyers.

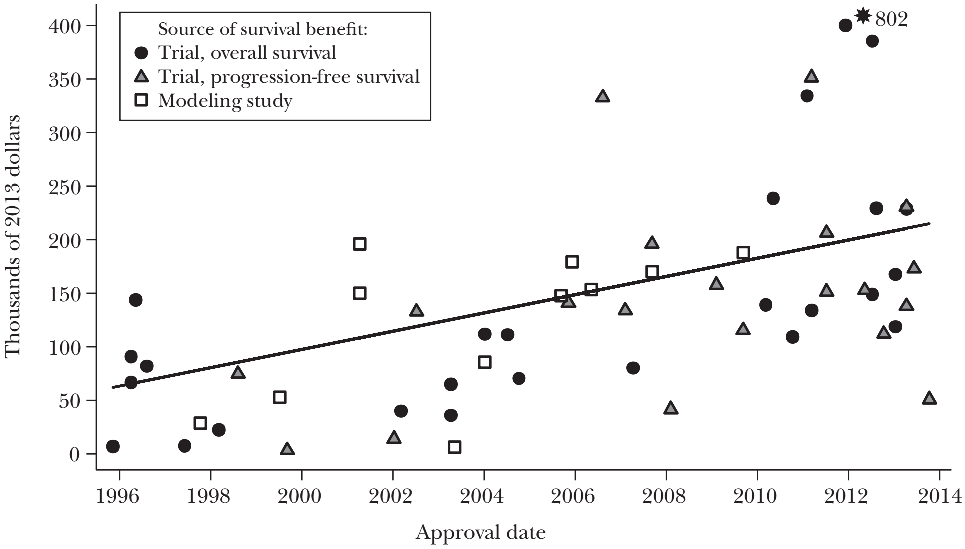
Our approach accounts for differences in the duration of treatment across drugs and is consistent with the notion of measuring the price of a treatment episode, as advocated by Berndt, Cutler, Frank, Griliches, Newhouse, and Triplett (2000) and Busch, Berndt, and Frank (2001). However, a drug's treatment episode price is not a comprehensive measure of the impact of that drug on health care costs. The impact of a drug on total costs depends on whether it is a substitute or complement to existing treatments and whether it increases or decreases the incidence of side effects, some of which can be quite costly to treat.

Prices versus Survival Benefits over Time

Figure 1 plots treatment-episode prices in 2013 dollars against incremental survival benefits, both on the natural log scale. The average drug price is $65,900 (in 2013 dollars), and the average survival benefit is 0.46 years. The markers identify drugs based on the source of survival benefit data: overall survival from a randomized trial; progression-free survival from a randomized trial; and overall survival from a modeling study. There is a positive correlation, 0.9, between treatment episode prices and incremental survival benefits. A regression of the natural logarithm of prices on incremental life-years gained indicates that prices increase by

Figure 2

Drug Price per Life Year Gained versus Drug Approval Date



Source: Authors.

Notes: The best fit line is: Price per life year gained = \$54,100 + \$8,500 × Approval Year. Approval Year = 0 for 1995, 1 for 1996, . . . 19 for 2014. For purposes of display, we recoded one value from \$802,000 to \$400,000.

120 percent (with a 95 percent confidence interval ranging from 74 to 166 percent) for each additional life-year gained (or 14 percent per month gained). The effect in dollar terms is \$75,000 per year gained (with a 95 percent confidence interval from \$12,000 to \$137,000).

Newer drugs are not associated with greater survival benefits compared to older drugs. A regression with life-years gained as the dependent variable and year of approval as the explanatory variable yields a small and insignificant coefficient (0.005 years of life gained, with a 95 percent confidence interval from −0.024 to 0.034 years of life gained).

Prices have increased over time. A regression of the natural logarithm of price on approval year indicates that prices increased by 12 percent per year (with a 95 percent confidence interval from 7 to 17 percent). The result is robust to the inclusion of a control for survival benefits.

For the remainder of the paper, we focus on trends in the price per life-year gained, which equals the price per treatment episode (in 2013 dollars) divided by survival benefits. The price per life-year gained can be thought of as a "benefit-adjusted" price. The sample average is \$150,100 per year of life gained (with a standard deviation of \$130,500). This value is in the range of estimates of the willingness-to-pay for a quality-adjusted life-year (Hirth, Chernow, Miller, Fendrick, and Weissert 2000). Figure 2 plots drugs' price per life-year gained against drugs' approval date. There is an upward trend. A regression of the price per life-year gained on approval year indicates that benefit- and inflation-adjusted launch prices increased by $8,500 (with a 95 percent confidence interval from $2,900 to $14,100) per year.[5] The intercept (1995 is zero on the x-axis) is $54,100 (95 percent confidence interval: −$16,700 to $124,900). Put another way, in 1995 patients and their insurers paid $54,100 for a year of life. A decade later, 2005, they paid $139,100 for the same benefit. By 2013, they were paying $207,000.

Figure 3 shows trends in the price per life-year classified by different types of anticancer drugs. Upward trends are apparent for most disease types.

Price Per Life-Year Gained and Drug Attributes

We used least squares regression to determine if the relationship between the price per life-year gained (in 2013 dollars) and approval year is robust to the inclusion of controls for other drug attributes. Table 1 presents regression estimates (sample means and other summary statistics for the drug attributes are presented in the Appendix available at http://e-jep.org). We used the natural logarithm of the price per life-year gained as the dependent variable because the price per life-year gained is skewed. Results are qualitatively similar if we use untransformed prices as the dependent variable. Because of the modest sample size, we did not attempt to control for all drug attributes simultaneously.

The model in column A, the baseline specification, indicates that benefit- and inflation-adjusted launch prices increased 10 percent per year over the study period.

The model in column B adds controls for the gastrointestinal complication and neutropenia rates. The gastrointestinal (GI) complication rate is the average of the nausea, vomiting, and diarrhea rates experienced by patients on the drug. The neutropenia rate is the proportion of patients who experience high-grade neutropenia, a deficit of white blood cells which puts patients at risk of infection. We set missing values to "0." Data on the side effects experienced by patients in the control arms of trials are inconsistently reported. We controlled for absolute rather than relative side effect rates, which may be why the coefficient on the gastrointestinal complication rate is "wrong signed." In general, side effect rates are similar for newer and older drugs (Niraula et al. 2012).

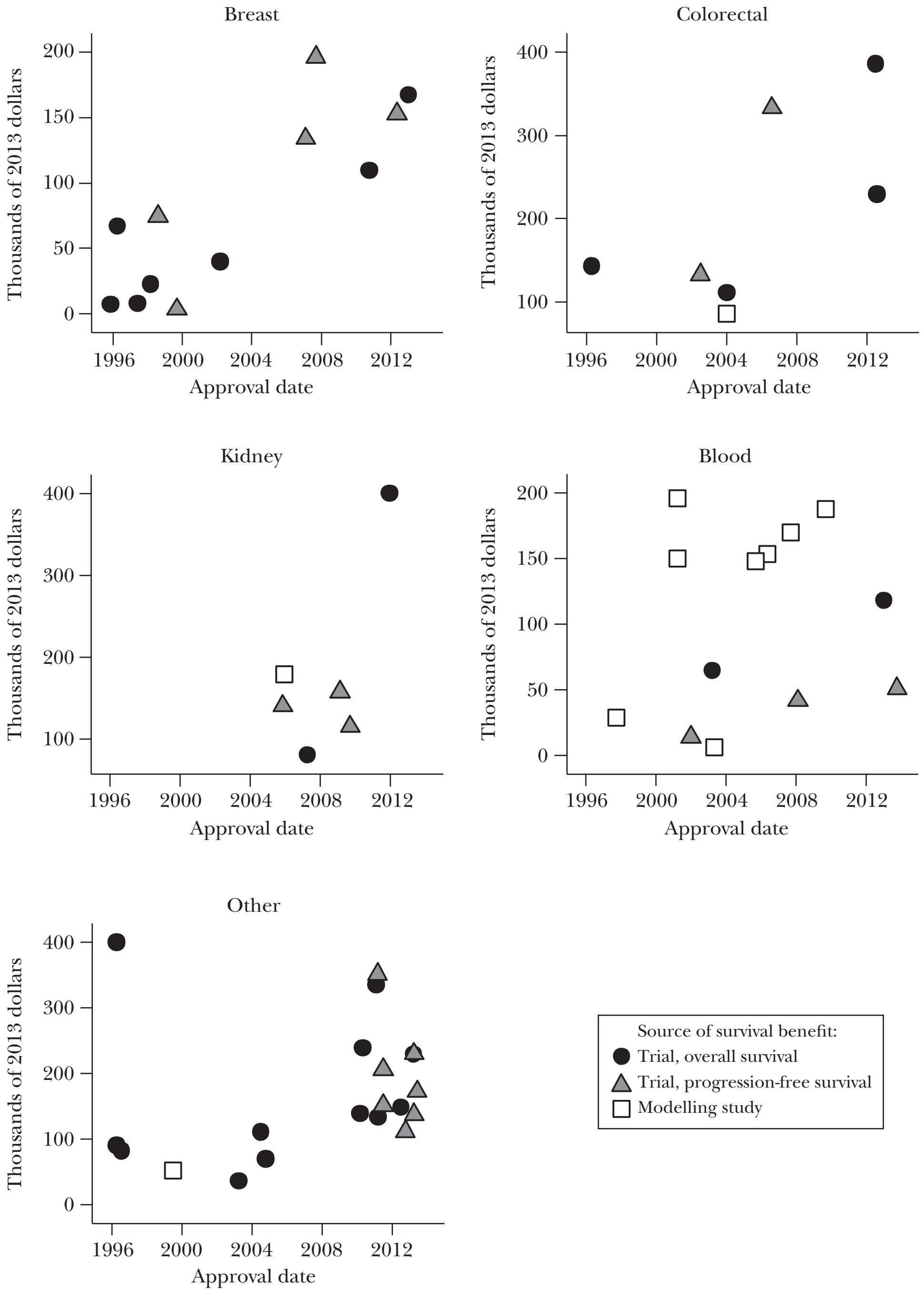
The model in column C includes a control for administration route: intravenous versus oral. Oral drugs are more convenient for patients than physician-administered intravenous drugs, but patients' out-of-pocket costs are typically higher for oral drugs. The positive coefficient on the intravenous administration route is insignificant.

The model in column D explores the hypothesis that increases in prices reflect increased production costs. We test this hypothesis indirectly by examining the link between several proxies for production costs and prices. Biologic drugs are typically more expensive to develop and produce than traditional anticancer drugs.

[5] The marginal effect from a generalized linear model with a log link and a gamma variance function is $8,500 (95 percent confidence interval: $1,800 to $15,300). Details of this approach are available in the online Appendix available with this paper at http://www.e-jep.org.

Figure 3

Drug Price per Life Year Gained versus Drug Approval Date by Indication



Source: Authors.

Table 1

Impact of Approval Year and Other Variables on the Natural Logarithm of the Price per Life Year Gained in 1,000s of 2013 US Dollars for 58 Cancer Drugs Approved between 1995 and 2013

	A	*B*	*C*	*D*	*E*	*F*
Approval year	0.10 [0.06, 0.14]*	0.10 [0.06, 0.14]*	0.10 [0.06, 0.14]*	0.10 [0.06, 0.15]*	0.10 [0.06, 0.15]*	0.09 [0.05, 0.13]*
GI complication rate		1.70 [0.47, 2.94]*				
Neutropenia rate		0.26 [−0.76, 1.28]				
IV drug			0.26 [−0.22, 0.74]			
Biologic				−0.15 [−0.67, 0.36]		
Multiproduct firm				0.38 [−0.14, 0.90]		
Randomized controlled trial					0.12 [−0.45, 0.69]	
Progression free survival					−0.36 [−0.91, 0.20]	
Placebo comparator						0.46 [−0.02, 0.94]+
Constant	3.51 [2.99, 4.03]*	2.95 [2.31, 3.59]*	3.34 [2.73, 3.95]*	3.24 [2.58, 3.89]*	3.48 [2.89, 4.06]*	3.39 [2.87, 3.92]*
R^2	0.28	0.37	0.29	0.31	0.30	0.32

	G	*H*	*I*	*J*	*K*
Approval year	0.10 [0.07, 0.14]*	0.10 [0.06, 0.14]*	0.09 [0.05, 0.14]*	0.09 [0.05, 0.13]*	0.11 [0.06, 0.15]*
Priority drug	0.93 [0.46, 1.40]*				
Orphan drug	−0.17 [−0.67, 0.33]				
Ln competitors		−0.64 [−0.99, −0.29]*			
Gene test			−0.59 [−1.05, −0.14]*		
Second line therapy			0.15 [−0.33, 0.62]		
Baseline survival				−0.29 [−0.53, −0.05]*	
Mortality rate					0.77 [−0.38, 1.92]
Constant	2.83 [2.23, 3.44]*	4.92 [4.01, 5.83]*	3.75 [3.09, 4.42]*	3.89 [3.30, 4.48]*	3.20 [2.50, 3.90]*
R^2	0.44	0.41	0.36	0.35	0.30

Notes: See text for definition of variables. 95 percent confidence intervals are in brackets. "GI" is gastrointestinal; "IV" is intravenous.

* Means significant at the 5 percent level, + means significant at the 10 percent level.

Multiproduct firms—firms that sell two or more anticancer drugs—are able to spread the fixed costs associated with marketing oncology drugs across products and may have equipment that can be used to manufacture two or more products. The coefficients on the cost-shifters are insignificant. These findings are consistent with the observation that there is a large gap between the generic and brand launch prices of anticancer drugs: for example, over 80 percent in the case of irinotecan (Conti et al. 2012). The prices of on-patent anticancer drugs do not appear to be closely related to marginal production costs.

The model in column E examines the relationship between the source of information about survival benefits and prices. We would expect that physicians would be more willing to prescribe drugs about which they have more information. This regression includes controls for whether the drug was approved on the basis of a randomized trial and if survival benefits are measured in terms of progression-free rather than overall survival. The coefficients are of the expected sign but are not significant.

The models in columns F–H consider whether drugs with few close substitutes command higher prices. Characterizing the degree of competition between anticancer drugs is difficult. Some compete, but most are used in a complementary manner, either in a co-administered multidrug "cocktail" regimen or in a sequence of therapy lines (first-line therapy, second-line therapy, etc.) Some drugs are approved to treat all patients diagnosed with late-stage cancer in a specific body part, while other drugs have narrower indications. The model in column F includes a control for whether the drug was compared against a placebo (or "best supportive care") or against another drug. Drugs tested against placebos occupy unique niches in the product space compared to drugs tested against "active" controls. Presumably the FDA and ethical review boards would not allow a manufacturer to test an anticancer drug against a placebo unless the drug had no direct substitutes. The coefficient is positive and significant at the 10 percent level. The model in column G includes controls for whether the drug was granted priority review status by the FDA. Priority review is granted to drugs that demonstrate "significant improvements in the safety or effectiveness of the treatment, diagnosis, or prevention of serious conditions when compared to standard applications." The model indicates that drugs awarded priority review status command significantly higher prices. The model in column H includes a variable equal to the natural logarithm of the number of drugs previously approved for the tumor site (National Cancer Institute 2014). The coefficient is negative and significant. It is unclear if this result can be interpreted as a purely competitive effect because anticancer drugs are often used in a complementary manner. The FDA grants orphan drug status to drugs used to treat rare conditions. The coefficient on orphan drug status (Model G) is not significant.

The model in column I includes controls for whether a drug was approved for use in patients with specific genetic biomarkers (US Food and Drug Administration 2014a) or as a second-line drug, for use in patients whose disease has progressed after an initial course of treatment. Demand may be less elastic, and prices higher, for drugs targeted at narrow patient subgroups. The coefficient on the gene test variable is negative, contrary to our expectation. The discussion up until this point has implicitly assumed that patients' valuation of gains in life expectancy from a new anticancer drug is independent of how long they could expect to live if they do not receive a new drug. This approach treats anticancer drugs as bundles of comparable attributes. The model in column J includes a control for baseline survival (as measured by survival in the control or comparator arm of the study we used to assess survival benefits). Results indicate that the longer patients survive without the drug, the lower the drug price. Patients' and physicians' willingness-to-pay may depend on absolute survival as well as relative survival gains. They may place a higher value on a drug that extends survival time by 6 months from a base of 8 months than one that extends survival time by 6 months from a base of 12 months.

The model in column K includes a control for the tumor-specific mortality rate, which we calculated by dividing the number of deaths attributed to the tumor by disease incidence. The coefficient on the mortality rate is positive but is not significant.[6]

The coefficient on approval year is economically and statistically significant in all 11 specifications in Table 1. Thus, our basic finding that benefit- and inflation-adjusted launch prices increased by about 10 percent annually appears robust to the inclusion of controls for the various drug attributes described above.

Sensitivity Checks

We performed several sensitivity checks. We re-estimated the baseline model (column A) on the subsample of drugs approved on the basis of randomized trials and for which we had trial-based estimates of overall survival. We also re-estimated the baseline model on the subsample of drugs with prices below the 90th percentile ($94,000) to determine the sensitivity of results to extreme values. In both cases the coefficients on approval date indicate that prices increased by 10 percent annually and were significant at the 1 percent level, consistent with the results from the baseline model.

Explaining Pricing Trends

Our empirical results suggest that the launch prices of anticancer drugs, even when adjusted for inflation and survival benefits, have increased substantially over time. We offer two explanations grounded in our observations of market behavior, economic theory, and current regulatory policy.

Our discussion focuses on the launch prices of branded drugs. If manufacturers make large changes to drugs' prices in the years following launch, our focus may be misplaced. We analyzed the Average Sales Price files from the Center for Medicare

[6] Mortality rates are measured with substantial error. Ideally, we would like to measure mortality among patients diagnosed with late-stage disease, but we do not have data on tumor incidence by stage at diagnosis.

and Medicaid Services for a subset of the drugs in our sample to determine if launch prices are a sufficient statistic for post-launch prices. The files capture prices for the mostly intravenous drugs reimbursed under Medicare's Part B outpatient medical benefit. We excluded three drugs—gemcitabine, irinotecan, and oxaliplatin—that experienced large declines in price following patent expiration and generic entry. We calculated annualized growth rates in the remaining sample of 19 drugs. The average annualized growth rate in real prices after launch was 1 percent. The 25th, 50th, and 75th percentiles were −0.7 percent, 0.9 percent, and 4 percent. The results are consistent with Lu and Comanor's (1998) finding that the prices of innovative drugs do not change much after launch. Launch prices are where the action is.

Reference Pricing

Writing to criticize the "astronomical" prices of new anticancer drugs, a group of over 100 prominent oncologists (Experts in Chronic Myeloid Leukemia 2013) proposed the following model of manufacturers' price setting behavior: "How are the prices of cancer drugs decided? Of the many complex factors involved, price often seems to follow a simple formula: start with the price for the most recent similar drug on the market and price the new one within 10–20 percent of that price (usually higher)." Industry insiders echo this theory of price-setting behavior. For example, from Hutchison (2010): "Gold [CEO of Dendreon] says that the cost of Provenge was based on the 'overall landscape' of treatment prices for cancer." From Marcus (2004): "A spokeswoman for AstraZeneca justified the price of Iressa as 'in line with other cancer treatments.'" From Silber (2005): "The retail price of the drug will be $5,416 per month, an amount that Onyx said is in the range of similarly specialized cancer drugs."

The theory that manufacturers set the prices of new drugs based on the prices of existing therapies (not necessarily competitors), rather than some intrinsic standard of product value, is consistent with reference price models of demand. Reference pricing models depart from the standard economic model of consumer behavior by allowing consumers' purchase decisions to depend on a pricing anchor, or reference price, rather than on an internal comparison of price and willingness-to-pay (Thaler 1985). Consumers may determine reference prices based on observed past prices or the prices of similar, but not necessarily substitute, goods.

Oncologists are in a strong position to influence the market share of anticancer drugs. Although oncologists do not face direct incentives to avoid costly drugs, they may balk at prescribing drugs with prices they perceive as exploitative—in the language of theory, drugs with prices above the reference price level. An extensive literature in economics and marketing describes how perceptions of fairness influence consumers' attitudes towards prices and market behavior (for example, Frey and Pommerehne 1993; Mas 2006; Maxwell 2002; Kahneman, Knetsch, and Thaler 1986; Piron and Fernandez 1995).

There is a "zone of indifference" around a reference price such that consumers ignore small deviations from the reference price (Kalyanaram and Little 1994). The zone of indifference gives manufacturers the ability to set the prices of new drugs slightly above the prices of existing drugs without reducing quantity demanded. As costlier drugs come to market, oncologists become habituated to higher prices, giving manufacturers leeway to set even higher prices in the future. The characteristics of the market for anticancer drugs, including patent protection, which protects producers from direct competition, and generous third party payment, allow this dynamic to persist. These characteristics are present in other medical product markets but not to the same degree as in the anticancer drug market.

Over time, the use of reference prices leads to forward-looking price complementarities between manufacturers. When a new drug enters with a price in excess of the reference price, it re-establishes price levels, freeing up the next entrant to set its price even higher. Kahneman, Knetsch, and Thaler (1986) write, "[P]rice increases that are not justified by increasing costs are judged less objectionable when competitors have led the way." Shortly after the FDA approved bevazicimab and erlotinib, one Wall St. analyst noted: "Companies will be looking at these products to help them determine the pricing of their own drugs . . . Tarceva and other drugs will likely take their cue from Erbitux and Avastin" (Griffith 2004). According to textbook monopoly pricing theory, the price of Erbitux (generic name cetuximab) should have had no direct bearing on the price of Tarceva (generic name erlotinib), a lung and pancreatic cancer drug, because cetuximab was not a competitor at the time.

If a manufacturer sets a price that is perceived as exploitative, in the sense that the price exceeds the reference price to a large degree, it risks provoking a backlash. One example of where this happened involved a second-line treatment for metastatic colorectal cancer, ziv-aflibercept (brand name Zaltrap). When approved by the FDA in 2012, its price was double that of bevacizumab, its closest competitor, at bevacizumab's common dosing level. Oncologists did not view ziv-aflibercept as particularly innovative, and three prominent physicians at the Memorial Sloan Kettering cancer center wrote an opinion piece in the *New York Times* (Bach, Saltz, and Wittes 2012) stating that they would refrain from using ziv-aflibercept at their center because of its price. One month later the manufacturer, Sanofi, announced that it would provide purchasers with a 50 percent discount off the list price.

According to one Wall Street analyst, "market structure effectively provides no mechanism for price control in oncology other than companies' goodwill and tolerance for adverse publicity" (Anand 2007). The observation begs the question: What is to stop a manufacturer from setting the price of a drug at $1,000,000 or more? Drug manufacturers are able to set higher prices for new drugs, but they must be mindful of physicians' ability to exact retribution when manufacturers violate physicians' norms of fairness in pricing.

Required Pricing Discounts

Recent increases in the launch prices of anticancer drugs may be an unintended consequence of policies to expand access to price discounts. The so-called 340B drug pricing program, authorized by Congress in 1992, requires drug manufacturers to provide deep discounts to 340B-qualified buyers. At the program's inception, only federally qualified health centers, specialized public health clinics, and "disproportionate share hospitals" (hospitals whose patient population includes a high proportion of low-income patients) qualified for 340B discounts. Discounts are set relative to the average price wholesalers, retail pharmacies, and providers pay manufacturers to purchase drugs, called the "Average Manufacturer Price." The 340B price discount for branded drugs must be at least 23.1 percent of the Average Manufacturer Price. Providers that purchase drugs through a government-designated distributor may receive additional discounts, though these are relatively small, totaling $67 million in 2013 (Drug Discount Monitor 2014). Participation in the 340B program is attractive for health care providers because they do not have to pass the discount on to insurers. They profit on the spread between third-party payers' drug reimbursement rates and the 340B discounted price.[7]

Since 1992, Congress and federal regulators have broadened eligibility to include critical access hospitals, free-standing cancer hospitals, some community hospitals, and outpatient clinics affiliated with disproportionate share hospitals. Mergers between 340B providers and non-340B providers, a predictable effect of the incentives inherent in the program, have also expanded the program's reach. Due to changes in eligibility rules and mergers, the number of providers in the 340B program increased from 8,605 in 2001 to 16,572 in 2011 (US General Accounting Office 2011). Industry sources predict that the volume of drug sales under the 340B program will increase from $6 billion in 2010 to $12 billion in 2016 (Biotechnology Industry Organization 2013).[8]

Because the 340B discount is based on a drug's average price, the program presents manufacturers with an incentive to set higher launch prices to offset discounts. Increases in the number of 340B-eligible providers have magnified the incentive, possibly leading to upward pressure in the prices paid by noneligible providers (Conti and Bach 2013). The 340B program also splits the market into price-elastic and price-inelastic segments. Just as branded drug manufacturers increase prices following generic entry to capture revenues from brand-loyal customers (Frank and Salkever 1997), manufacturers of recently launched drugs may cede large discounts to their price-sensitive segment but increase prices to non-340B providers.

The federal Medicaid program has its own set of drug pricing rules. In exchange for formulary coverage by state Medicaid programs, branded manufacturers give rebates to the federal government on sales to Medicaid patients. Similar to the 340B program, the rebate is based on the Average Manufacturer Price. If a manufacturer increases the price of a drug over and above the rate of inflation, it must pay a larger rebate. This aspect of the program provides incentives for firms to set higher prices initially, rather than increasing prices after launch. Although Medicaid

[7] When calculating average sales prices for purposes of Medicare reimbursement, regulations instruct manufacturers to exclude sales to 340B providers. Hence Medicare reimbursement rates are not affected by growth in the 340B discount program, though providers' acquisition costs are reduced.

[8] This figure includes anticancer and noncancer drugs. Industry sources indicate that the two therapeutic classes having the largest 340B sales are anticancer drugs and anti-infectives.

accounts for less than 10 percent of spending on cancer treatment (Howard, Molinari, and Thorpe 2004), enrollment in the program is growing, presenting manufacturers with additional incentives to increase prices to non-Medicaid patients.

The United Kingdom and other European countries negotiate drug prices with manufacturers. Although negotiated discounts are not legislatively linked to the US price, the US price may serve as an opening bid in negotiations, and discounts are often expressed as a percent of the US list price in contracts. As pressure has mounted on governments to reign in health spending, European health systems have adopted a more aggressive bargaining stance, backed by a credible threat of noncoverage, potentially leading manufacturers to set higher US prices.[9] The United Kingdom and many other countries do not divulge negotiated drug prices, and so we are unable to determine whether launch prices have increased outside the United States. There is anecdotal evidence that they have. For example, a number of signatories to a statement calling attention to the "unsustainable" prices of new anticancer drugs were European physicians (Experts in Chronic Myeloid Leukemia 2013).

Other Potential Causes of Price Increases

What about other possible explanations for pricing trends, such as shifts in patient or physician demand? Changes on the demand side of the market seem inconsistent with observed pricing trends. The income elasticity of the demand for health care is not large enough to account for changes in prices or health care spending generally (Newhouse 1992). Moreover, patient cost-sharing is higher now than it was in 1995 as consumers have shifted to high-deductible plans (Berndt and Newhouse 2012; Kaiser Family Foundation 2013). The structure of insurers' payments to physicians has remained largely unchanged, but payment levels for physician-administered anticancer drugs have declined following passage of the Medicare Modernization Act in 2003 (Jacobson et al. 2006; Jacobson, Earle, Price, and Newhouse 2010).

On the supply side, it is unlikely that changes in development and production costs alone can explain launch pricing trends. The FDA has reduced barriers to approval, and advances in genetics have facilitated drug discovery. The generic versions of anticancer drugs cost much less than the branded versions, suggesting that production costs are low relative to pre-patent expiration price levels. Pharmaceutical manufactures often claim that they set drug prices to recoup research and development costs. Manufacturers' research and development costs may have increased over time. As more drugs come to market, the number of unexploited targets for anticancer therapy shrinks, requiring firms to invest more to develop new drugs. Lacking measures of research and development costs, we are unable to evaluate the claim empirically. However, research and development costs are sunk

[9] The British National Health Service and other national health systems do not disclose negotiated prices, and so we cannot determine whether the spread between domestic and international drug prices has increased.

at the time of product launch and so they ought not to factor into the pricing decisions of a profit-maximizing firm once the product has been developed. We believe the direction of causation runs from prices to research and development costs—as prices increase, manufacturers are willing to spend more to discover new drugs—rather than the other way around.

Discussion

We find that, controlling for inflation and survival benefits, the launch prices of new anticancer drugs have increased over time. We do not anticipate that US payers and providers will change their policies in a way that will fundamentally change pricing dynamics, at least in the near term. The American Society of Clinical Oncology, the main professional group for physicians who treat cancer patients, is encouraging its members to consider costs when they choose drugs, but these efforts are mostly focused on costs to patients rather than systemwide costs. Efforts to increase the sensitivity of physician demand to drug prices still rely on physicians' sense of fairness rather than their pocketbooks. A Congressional advisory board, the Medicare Payment Advisory Commission, recently held a hearing on reforming reimbursement for physician-administered drugs. Many committee members voiced support for proposals that would reduce Medicare reimbursement for drugs if there are less-costly alternatives that have a "similar health effect" (InsideHealthPolicy 2014). However, newly-approved anticancer drugs are, by definition, unique, and will probably be unaffected if Medicare implements the policy.

To supporters of the US health care system, new anticancer drugs are a potent symbol of progress and represent the type of innovation that would be squelched if Medicare and other US insurers denied coverage to costly treatments (for example, Gingrich 2009). To critics, the pricing of new anticancer drugs represents the worst excesses of a system that provides few checks on drug companies' pricing power and prioritizes gains in health, however small, over cost control. Policymakers are quick to agree that the health system should discourage use of ineffective treatments, but it is unclear how regulators, insurers, and physicians should approach treatments that are more costly but also offer small incremental benefits.

The optimistic view of recent trends in cancer drug development is that although individual drugs may not be associated with large gains in survival, the work that goes into developing a new drug contributes to the stock of knowledge about cancer biology. Eventually, scientists will use the information gleaned from the development of existing drugs to develop new drugs with much greater benefits. The pessimistic view is that current coverage, reimbursement, and patent policies (Budish, Roin, and Williams 2013) divert drug manufacturers' attention away from developing drugs that yield truly meaningful survival benefits. If insurers restricted coverage to drugs that improved survival time by an economically significant amount, perhaps there would be more of them.

■ *Rena M. Conti received funding from a K07 CA138906 award from the National Cancer Institute to the University of Chicago.*

References

Akscin, John, Thomas R. Barr, and Elaine L. Towle. 2007. "Key Practice Indicators in Office-Based Oncology Practices: 2007 Report on 2006 Data." *Journal of Clinical Oncology* 3(4): 200–203.

Anand, Geeta. 2007. "Prescribing Caution: From Wall Street, a Warning About Cancer-Drug Prices; Morgan Stanley Analyst Creates Stir in Industry as He Sees a Backlash." *Wall Street Journal,* March 15. http://online.wsj.com/news/articles/SB117391934158537592.

Bach, Peter B. 2009. "Limits on Medicare's Ability to Control Rising Spending on Cancer Drugs." *New England Journal of Medicine* 360(6): 626–33.

Bach, Peter B., Leonard B. Saltz, and Robert E. Wittes. 2012. "In Cancer Care, Cost Matters." *New York Times,* October 14.

Berndt, Ernst R., Linda Bui, David R. Reiley, and Glen L. Urban. 1995. "Information, Marketing, and Pricing in the U.S. Antiulcer Drug Market." *American Economic Review* 85(2): 100–105.

Berndt, Ernst R., David M. Cutler, Richard G. Frank, Zvi Griliches, Joseph P. Newhouse, and Jack E. Triplett. 2000. "Medical Care Prices and Output." Chap. 3 in *Handbook of Health Economics,* Vol. 1A, edited by Anthony C. Culyer and Joseph P. Newhouse. Amsterdam: Elsevier Science.

Berndt, Ernst R., and Joseph P. Newhouse. 2012. "Pricing and Reimbursement in US Pharmaceutical Markets." Chap. 8 in *The Oxford Handbook of the Economics of the Biopharmaceutical Industry,* edited by Patricia M. Danzon and Sean Nicholson. Oxford University Press.

Biotechnology Industry Organization. 2013. *The 340B Drug Discount Program.* https://www.bio.org/articles/340b-drug-discount-program.

Budish, Eric, Benjamin N. Roin, and Heidi Williams. 2013. "Do Fixed Patent Terms Distort Innovation? Evidence from Cancer Clinical Trials." NBER Working Paper 19430.

Busch, Susan H., Ernst R. Berndt, and Richard G. Frank. 2001. "Creating Price Indexes for Measuring Productivity in Mental Health Care." Chap. 5 in *Frontiers in Health Policy Research,* vol. 4, edited by Alan M. Garber. Cambridge, MA: MIT Press for the National Bureau of Economic Research.

Cavalli, Franco. 2013. "An Appeal to World Leaders: Stop Cancer Now." *The Lancet* 381(9865): 425–26.

Caves, Richard E., Michael D. Whinston, and Mark A. Hurwitz. 1991. "Patent Expiration, Entry, and Competition in the U.S. Pharmaceutical Industry." *Brookings Papers on Economic Activity: Microeconomics* 1–48.

Center for Medicare & Medicaid Services. 2014. "CMS Proposes Program Changes for Medicare Advantage and Prescription Drug Benefit Programs for Contract Year 2015 (CMS-4159-P)." January 6. http://www.cms.gov/Newsroom/MediaReleaseDatabase/Fact-Sheets/2014-Fact-sheets-items/2014-01-06.html.

CenterWatch. 2014. FDA Approved Drugs for Oncology. http://www.centerwatch.com/drug-information/fda-approved-drugs/therapeutic-area/12/.

Chen, Jie, and John A. Rizzo. 2012. "Pricing Dynamics and Product Quality: The Case of Antidepressant Drugs." *Empirical Economics* 42(1): 279–300.

Cockburn, Iain M., and Aslam H. Anis. 2001. "Hedonic Analysis of Arthritis Drugs." Chap. 11 in *Medical Care Output and Productivity,* edited by David Cutler and Ernst R. Berndt. University of Chicago Press.

Conti, Rena M., and Peter B. Bach. 2013. "Cost Consequences of the 340B Drug Discount Program." *JAMA: Journal of the American Medical Association* 309(19): 1995–96.

Conti, Rena M., Arielle C. Bernstein, Victoria M. Villaflor, Richard L. Schilsky, Meredith B. Rosenthal, and Peter B. Bach. 2013. "Prevalence of Off-Label Use and Spending in 2010 Among

Patent-Protected Chemotherapies in a Population-Based Cohort of Medical Oncologists." *Journal of Clinical Oncology* 31(9): 1134–39.

Conti, Rena M., Meredith B. Rosenthal, Blase N. Polite, Peter B. Bach, and Ya-Chen Tina Shih. 2012. "Infused Chemotherapy Use in the Elderly After Patent Expiration." *American Journal of Managed Care* 18(5): e173–78.

Dendreon. 2014. Patient Access. http://www.dendreon.com/patient_resources/patient_access/.

Drug Discount Monitor. 2014. "340B Sales Totaled $7.5 Billion in 2013, Apexus Says." February 14. Link (for subscribers only): http://drugdiscountmonitor.com/2014/02/340b-sales-totaled-7-5-billion-in-2013-apexus-says/.

Experts in Chronic Myeloid Leukemia. 2013. "The Price of Drugs for Chronic Myeloid Leukemia (CML) is a Reflection of the Unsustainable Prices of Cancer Drugs: From the Perspective of a Large Group of CML Experts." *Blood* 121(22): 4439–42.

Fleck, Leonard M. 2013. "Just Caring: Can We Afford the Ethical and Economic Costs of Circumventing Cancer Drug Resistance?" *Journal of Personalized Medicine* 3(3): 124–43.

Frank, Richard G., and David S. Salkever. 1997. "Generic Entry and the Pricing of Pharmaceuticals." *Journal of Economics and Management Strategy* 6(1): 75–90.

Frey, Bruno S., and Werner W. Pommerehne. 1993. "On the Fairness of Pricing—An Empirical Survey among the General Population." *Journal of Economic Behavior and Organization* 20(3): 295–307.

Gingrich, Newt. 2009. "Trust the Government." *Human Events*, August 12. http://humanevents.com/2009/08/12/trust-the-government/.

Goldman, Dana P., Anupam B. Jena, Darius N. Lakdawalla, Jennifer L. Malin, Jesse D. Malkin, and Eric Sun. 2010. "The Value of Specialty Oncology Drugs." *Health Service Research* 45(1): 115–32.

Goldman, Dana P., Geoffrey F. Joyce, Grant Lawless, William H. Crown, and Vincent Willey. 2006. "Benefit Design and Specialty Drug Use." *Health Affairs* 25(5): 1319–31.

Griffith, Victoria. 2004. "Pricing Weighs on Cancer Treatments." *Financial Times,* June 3.

Hall, Stephen S. 2013. "The Cost of Living." *New York Magazine*, October 20. http://nymag.com/news/features/cancer-drugs-2013-10/.

Hirth, Richard A, Michael E. Chernew, Edward Miller, A. Mark Fendrick, and William G. Weissert. 2000. "Willingness to Pay for a Quality-Adjusted Life Year: In Search of a Standard." *Medical Decision Making* 20(3): 332–42.

Howard, David H. 2014. "Drug Companies' Patient-Assistance Programs—Helping Patients or Profits?" *New England Journal of Medicine* 371(2): 97–99.

Howard, David H., John Kauh, and Joseph Lipscomb. 2010. "The Value of New Chemotherapeutic Agents for Metastatic Colorectal Cancer." *Archives of Internal Medicine* 170(6): 537–42.

Howard, David H., Carolyn Kenline, Hillard M. Lazarus, Charles F. LeMaistre, Richard T. Maziarz, Philip L. McCarthy Jr., Susan K. Parsons, David Szwajcer, James Douglas Rizzo, and Navneet S. Majhail. 2011. "Abandonment of High Dose Chemotherapy/Hematopoietic Cell Transplants for Breast Cancer." *Health Services Research* 46(6, Part 1): 1762–77.

Howard, David H., Nicole A. Molinari, and Kenneth E. Thorpe. 2004. "National Estimates of Medical Costs Incurred by Nonelderly Cancer Patients." *Cancer* 100(5): 883–91.

Hutchison, Courtney. 2010. "Provenge Cancer Vaccine: Can You Put a Price on Delaying Death?" *ABC News*, July 29. http://abcnews.go.com/Health/ProstateCancerNews/provenge-cancer-vaccine-months-life-worth-100k/story?id=11269159.

IMS Institute for Healthcare Informatics. 2014. "Innovation in Cancer Care and Implications for Health Systems: Global Oncology Trend Report." Available at: http://www.imshealth.com/portal/site/imshealth/menuitem.762a961826aad98f53c753c71ad8c22a/?vgnextoid=f8d4df7a5e8b5410VgnVCM10000076192ca2RCRD.

InsideHealthPolicy. 2014. "MedPAC Explores Fixes to Part B Drug Payment Policy, Calls Current System 'Perverse Incentive.'" November 12. http://www.pipcpatients.org/pipc-admin/pdf/cde114_MedPAC%20Explores%20Fixes%20To%20Part%20B%20Drug%20Payment%20Policy.pdf.

Jacobson, Mireille, Craig C. Earle, Mary Price, and Joseph P. Newhouse. 2010. "How Medicare's Payment Cuts for Cancer Chemotherapy Drugs Changed Patterns of Treatment." *Health Affairs* 29(7): 1391–99.

Jacobson, Mireille, A. James O'Malley, Craig C. Earle, Juliana Pakes, Peter Gaccione, and Joseph P. Newhouse. 2006. "Does Reimbursement Influence Chemotherapy Treatment for Cancer Patients?" *Health Affairs* 25(2): 437–43.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review* 76(4): 728–41.

Kaiser Family Foundation and Health Research & Educational Trust. 2013. *Employer Health Benefits: 2013 Annual Survey.* Washington, DC.

Kalyanaram, Gurumurthy, and John D. C. Little. 1994. "An Empirical Analysis of Latitude of Price

Acceptance in Consumer Package Goods." *Journal of Consumer Research* 21(3): 408–18.

Kantarjian, Hagop M., Tito Fojo, Michael Mathisen, and Leonard A. Zwelling. 2013. "Cancer Drugs in the United States: *Justum Pretium*—The Just Price." *Journal of Clinical Oncology* 31(28): 3600–3604.

Lakdawalla, Darius N., John A. Romley, Yuri Sanchez, J. Ross Maclean, John R. Penrod, and Tomas Philipson. 2012. "How Cancer Patients Value Hope and the Implications for Cost-Effectiveness Assessments of High-Cost Cancer Therapies." *Health Affairs* 31(4): 676–82.

Landsem, Hope. 2013. "Rationing Health Care in Oregon." *Wall Street Journal,* August 8. http://online.wsj.com/news/articles/SB10001424127887324522504579000560184822956.

Lichtenberg, Frank R. 2009a. "The Effect of New Cancer Drug Approvals on the Life Expectancy of American Cancer Patients, 1978–2004." *Economics of Innovation and New Technology* 18(5): 407–28.

Lichtenberg, Frank R. 2009b. "International Differences in Cancer Survival Rates: The Role of New Drug Launches." *International Journal of Healthcare Technology and Management* 10(3): 138–55.

Lu, Z. John, and William S. Comanor. 1998. "Strategic Pricing of New Pharmaceuticals." *Review of Economics and Statistics* 80(1): 108–18.

Lucarelli, Claudio, and Sean Nicholson. 2009. "A Quality-Adjusted Price Index for Colorectal Cancer Drugs." NBER Working Paper 15174.

Malin, Jennifer L., Jane C. Weeks, Arnold L Potosky, Mark C. Hornbrook, and Nancy L. Keating. 2013. "Medical Oncologists' Perceptions of Financial Incentives in Cancer Care." *Journal of Clinical Oncology* 31(5): 530–35.

Marcus, Amy Dockser. 2004. "Price Becomes Factor in Cancer Treatment." *Wall Street Journal,* September 7. http://online.wsj.com/news/articles/SB109450779986210547.

Mas, Alexandre. 2006. "Pay, Reference Points, and Police Performance." *Quarterly Journal of Economics* 121(3): 783–821.

Maxwell, Sarah. 2002. "Rule-Based Price Fairness and Its Effect on Willingness to Purchase." *Journal of Economic Psychology* 23(2): 191–212.

National Cancer Institute. 2014. "Cancer Drug Information." A webpage: http://www.cancer.gov/cancertopics/druginfo/alphalist.

National Comprehensive Cancer Network. 2010. "Guidelines for Breast Cancer Updated; Bevacizumab Recommendation Affirmed." October 18. http://www.nccn.org/about/news/newsinfo.aspx?NewsID=259.

Neumann, Peter J. 2005. *Using Cost-Effectiveness Analysis to Improve Health Care: Opportunities and Barriers.* Oxford University Press.

Newhouse, Joseph P. 1992. "Medical Care Costs: How Much Welfare Loss?" *Journal of Economic Perspectives* 6(3): 3–21.

Niraula, Saroj, Bostjan Seruga, Alberto Ocana, Tiffany Shao, Robyn Goldstein, Ian F. Tannock, and Eitan Amir. 2012. "The Price We Pay for Progress: A Meta-Analysis of Harms of Newly Approved Anticancer Drugs." *Journal of Clinical Oncology* 30(24): 3012–19.

PhRMA (Pharmaceutical Research and Manufacturers of America). 2014. "Medicines in Development." A webpage: http://www.phrma.org/innovation/meds-in-development.

Piron, Robert, and Luis Fernandez. 1995. "Are Fairness Constraints on Profit-Seeking Important?" *Journal of Economic Psychology* 16(1): 73–96.

Raborn Martin L., Elise M. Pelletier, Daniel B. Smith, and Carolina M. Reyes. 2012. "Patient Out-of-Pocket Payments for Oral Oncolytics: Results from a 2009 US Claims Data Analysis." *American Journal of Managed Care* 18(5 No. 2): SP57–64.

Reekie, W. Duncan. 1978. "Price and Quality Competition in the United States Drug Industry." *Journal of Industrial Economics* 26(3): 223–37.

Romley, John A, Yuri Sanchez, John R. Penrod, and Dana P. Goldman. 2012. "Survey Results Show That Adults Are Willing to Pay Higher Insurance Premiums for Generous Coverage of Specialty Drugs." *Health Affairs* 31(4): 683–90.

Schrag, Deborah. 2004. "The Price Tag on Progress—Chemotherapy for Colorectal Cancer." *New England Journal of Medicine* 351(4): 317–19.

Seabury, Seth A., Dana P. Goldman, J. Ross Maclean, John R. Penrod, and Darius N. Lakdawalla. 2012. "Patients Value Metastatic Cancer Therapy More Highly Than is Typically Shown Through Traditional Estimates." *Health Affairs* 31(4): 691–99.

Silber, Judy. 2005. "Onyx Gets OK for Kidney Cancer Drug." *Contra Costa Times,* December 21.

Snider, Julia Thornton, John A. Romley, William B. Vogt, and Tomas J. Philipson. 2012. "The Option Value of Innovation." *Forum for Health Economics & Policy* 15(2).

Sun, Eric, Anupam B. Jena, Darius Lakdawalla, Carolina Reyes, Tomas J. Philipson, and Dana Goldman. 2010. "The Contributions of Improved Therapy and Early Detection to Cancer Survival Gains, 1988–2000." *Forum for Health Economics & Policy* 13(2).

Suslow, Valerie Y. 1996. "Measuring Quality Change in Pharmaceutical Markets: Hedonic Price Indexes for Anti-Ulcer Drugs." Chap. 4 in *Competitive Strategies in the Pharmaceutical Industry,* edited by Robert B. Helms. Washington, DC: American Enterprise Institute.

Szabo, Liz. 2004 "Price of Cancer Drugs Called

'Mind-Boggling.'" *USA Today*, July 21. http://usatoday30.usatoday.com/news/health/2004-07-21-cancer-usat_x.htm.

Thaler, Richard. 1985. "Mental Accounting and Consumer Choice." *Marketing Science* 4(3): 199–214.

Trusheim, Mark R., and Ernst R. Berndt. 2012. "The Segmentation of Therapeutic Populations in Oncology." *Health Management, Policy and Innovation* 1(1): 19–34.

US Food and Drug Administration. 2014a. Table of Pharmacogenomic Biomarkers in Drug Labeling. http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm.

US Food and Drug Administration. 2014b. "Fast Track, Breakthrough Therapy, Accelerated Approval, Priority Review." http://www.fda.gov/ForPatients/Approvals/Fast/default.htm.

US General Accounting Office. 2011. *Manufacturer Discounts in the 340B Program Offer Benefits, But Federal Oversight Needs Improvement.* Washington, D.C.

Weeks, Jane C., Paul J. Catalano, Angel Cronin, Matthew D. Finkelman, Jennifer W. Mack, Nancy L. Keating, and Deborah Schrag. 2012 "Patients' Expectations about Effects of Chemotherapy for Advanced Cancer." *New England Journal of Medicine* 367(17): 1616–25.

Wittes, Robert E. 2004. "Cancer Weapons, Out of Reach." *Washington Post.* June 15. http://www.washingtonpost.com/wp-dyn/articles/A42035-2004Jun14.html.

Woodward, Rebecca M., Martin L. Brown, Susan T. Stewart, Kathleen A. Cronin, and David M. Cutler. 2007. "The Value of Medical Interventions for Lung Cancer in the Elderly: Results from SEER-CMHSF." *Cancer* 110(11): 2511–18.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 163–180

The Window Tax: A Case Study in Excess Burden[†]

Wallace E. Oates and Robert M. Schwab

> "*The adage 'free as air' has become obsolete by Act of Parliament. Neither air nor light have been free since the imposition of the window-tax. We are obliged to pay for what nature lavishly supplies to all, at so much per window per year; and the poor who cannot afford the expense are stinted in two of the most urgent necessities of life.*"
>
> — Charles Dickens (1850, p. 461)

The window tax provides a dramatic and transparent historical example of the potential distorting effects of taxation. Imposed in England in 1696, the tax—a kind of predecessor of the modern property tax—was levied on dwellings with the tax liability based on the number of windows. The tax led to efforts to reduce tax bills through such measures as the boarding up of windows and the construction of houses with very few windows. Sometimes whole floors of houses were windowless. In spite of the pernicious health and aesthetic effects and despite widespread protests, the tax persisted for over a century and a half: it was finally repealed in 1851.

Our purpose in this paper is threefold. First, we provide a brief history of the tax with a discussion of its rationale, its role in the British fiscal system, and its economic and political ramifications. Second, we have assembled a dataset from microfilms of local tax records during this period that indicate the numbers of

■ *Wallace E. Oates and Robert M. Schwab are, respectively, Distinguished University Professor Emeritus and Professor of Economics, Department of Economics, University of Maryland, College Park, Maryland. Their email addresses are oates@econ.umd.edu and schwab@econ.umd.edu.*

[†] To access the Appendix, Data Appendix, and disclosure statements, visit
http://dx.doi.org/10.1257/jep.29.1.163

doi=10.1257/jep.29.1.163

windows in individual dwellings. Drawing on these data, we are able to test some basic hypotheses concerning the effect of the tax on the number of windows and to calculate an admittedly rough measure of the excess burden associated with the window tax. Third, we have in mind a pedagogical objective. The concept of excess burden (or "deadweight loss") is for economists part of the meat and potatoes of tax analysis. But to the laity the notion is actually rather arcane; public-finance economists often have some difficulty, for example, in explaining to taxpayers the welfare costs of tax-induced distortions in resource allocation. The window tax is a textbook example of how a tax can have serious adverse side effects on social welfare.[1] In addition to its objectionable consequences for tax equity, the window tax resulted in obvious and costly misallocations of resources.

A Brief History of the Window Tax

The window tax was introduced in England in 1696 by King William III.[2] Burdened with expenses from the Revolution, the war with France, and the costs of re-coinage necessitated by the "miserable state" of existing coins, which had been reduced by "clipping" (the scraping-off of small portions of the high-grade silver coins), the King levied a new tax consisting originally of a flat rate of 2 shillings upon each house and an additional charge of 4 shillings upon houses with between 10 and 20 windows and 8 shillings upon houses with more than 20 windows (Dowell 1965, vol. 3, p. 168). The tax was intended to be a temporary levy, but it was restructured and increased several times. In the end, the window tax lasted in various forms for over 150 years; as we noted above, it was not repealed until 1851.

An important feature of the tax was that it was levied on the occupant, not the owner of the dwelling. Thus, the renter, not the landlord, paid the tax. However, large tenement buildings in the cities, each with several apartments, were an exception. They were charged as single residences with the tax liability resting on the landlord. This led to especially wretched conditions for the poor in the cities, as landlords blocked up windows and constructed tenements without adequate light and ventilation (Glantz 2008, p. 33).

Although the rate structure of the window tax was revised numerous times over this lengthy period, one feature is of special importance for our study. The tax did not consist of a series of smoothly rising marginal rates but instead included a series of "notches"—points at which an additional window brought with it a large

[1] And in fact, several textbooks offer the window tax as an example of a tax that distorts economic decisions. See, for example, Stiglitz (1988, p. 17), Mateer and Coppock (2014, p. 201), and Rosen and Gayer (2010, p. 369).

[2] This section draws heavily on Glantz (2008), who provides by far the most careful and thorough treatment of the history of the window tax. For other treatments of the tax, see Ward (1952), Beckett (1985), and Dowell (1965, vol. 3, pp. 168–192; first published in 1884). For useful histories of taxation in the United Kingdom that cover this period and address the window tax, see Sinclair (1803), Dowell (1884, vol. 3), Kennedy (1913), Binney (1958), and Douglas (1999).

increase in tax liability. Consider, for example, the reforms introduced in 1747, under which Parliament raised and recast the rate structure of the tax. The fixed 2 shillings per dwelling was detached from the window tax and imposed in addition to a new schedule of rates of windows. Under the new rate schedule, there was a tax of 6 pence *on every window* in a house with 10 to 14 windows, of 9 pence per window in houses with 15 to 19 windows, and of 1 shilling for every window in houses with more than 20 windows.[3] As a result, we might expect to find, for example, many more houses with 9, rather than 10, windows. We will make use of these notches in our later empirical study of the effects of the tax.

The window tax, incidentally, had an antecedent: the hearth tax. Imposed in 1662 by Charles II after the Restoration, the hearth tax consisted of a levy of 2 shillings for every fire-hearth and stove in houses in England and Wales. The tax was very unpopular in part because of the intrusive character of the assessment process. The "chimney-men" (as the assessors and tax collectors were called) had to enter the house to count the number of hearths and stoves, and there was great resentment against this invasion of the sanctity of the home. The window tax, in contrast, did not require access to the interior of the dwelling: the "window peepers" could count windows from the outside, thus simplifying the assessment procedure and obviating the need for an invasion of the interior.

Both of these taxes were intended to be a visible indicator of ability to pay. As pointed out in a discussion in the House of Commons (1850) just prior to the repeal of the window tax, "The window tax, when first laid on, was not intended as a window tax, but as a property tax, as a house was considered a safe criterion of the value of a man's property, and the windows were only assumed as the index of the value of houses." But as Adam Smith (1776 [1937], p. 798) observed in *The Wealth of Nations*, the number of windows could be a very poor measure of the value of a dwelling: "A house of ten pounds rent in the country may have more windows than a house of five hundred pounds rent in London; and though the inhabitant of the former is likely to be a much poorer man than that of the latter, yet so far as his contribution is regulated by the window-tax, he must contribute more to the support of the state."

Although the window tax removed the need for tax assessors to enter the house to count the number of hearths, the tax created some administrative problems of its own—not the least of which was the definition of a "window" for purposes of taxation. In 1848, for example, Professor Scholefield of Cambridge paid tax on a hole in the wall of his coal cellar (House of Commons 1848). In the same year, Mr. Gregory Gragoe of Westminster paid tax for a trapdoor to his cellar (House of Commons 1848). An individual might have to pay tax should a brick fall out of the wall if the hole admitted light into the house. Indeed, if the dwelling was already at one of the "notch" points for the tax, a new hole from a missing brick could force the resident to pay a higher rate on every window in the house. This issue was a source of considerable unrest

[3] There were 20 shillings to the pound and 12 pence to the shilling. The average annual income during this period was a bit less than 20 pounds per year.

among taxpayers. As late as 1850, there were continued requests to the Chancellor of the Exchequer for clarifications on the definition of a window.

The schedule and levels of rates for the window tax were amended (in some cases raised dramatically) over the life of the tax. As mentioned earlier, the original rate structure in 1696 was recast in 1747. Then in 1761, a tax rate of 1 shilling per window was established for homes with 8 or 9 windows and rates were raised on homes with 10 or more windows. We look at the effect of the 1761 tax rate changes in a later section of the paper.

Significant changes continued to be made before the tax was eventually repealed. In 1784, Prime Minister William Pitt increased tax rates to compensate for lower taxes on tea. In response, "Owners in both town and country began to disfigure their houses . . . by blocking up their windows" (House of Commons 1848). In 1797, Pitt's Triple Assessment Act tripled the window tax rates to help pay for the Napoleonic Wars. The day following this new Act, thousands of windows were blocked up, and "Lighten our darkness we beseech thee, O Pitt!" was written in chalk on the blocked-up spaces (House of Commons 1848). There were some reductions in the window tax after 1820.

There were some exemptions under the window tax. Various factories and buildings were exempted from the tax: public offices, farm houses that cost less than 200 pounds per year, dairies, cheese rooms, malt houses, granaries, and coach makers. The rationale for these exemptions was either of two conditions: the windows provided air rather than light, or the trade required ample light so that workshops had to have glass windows. In addition, officials exempted some residences under various pretexts. Some exceptions were made for certain wealthy parties. In some instances, the presence of serious disease resulted in tax exemption. As stated in a decree in 1819, "In cases where the terror of contagion had forced the wretched inhabitants to restore the windows, and admit the light and air, the tax so incurred should be remitted" (House of Commons 1819). Such exemptions were a source of considerable controversy.

England and Scotland were both subject to the window tax, but Ireland was exempted because of its impoverished state. Some members of Parliament joked: "In advocating the extension of the window-tax to Ireland, the hon. Gentleman seemed to forget the fact that an English window and an Irish window were very different things. In England, the window was intended to let the light in; but in Ireland the use of a window was to let the smoke out" (House of Commons 1819).

The Adverse Health and Aesthetic Effects of the Window Tax

Much of the controversy over the window tax involved its highly regressive incidence, and the tax did indeed burden the poor.[4] However, the distorting effects

[4] In Appendix G to his *Principles of Economics*, Alfred Marshall (1890 [1948]] discusses the window tax in a footnote. Like Smith, his concern is solely with the incidence of the tax (not with its effects

on resource allocation were perhaps even more widespread and profound. Residents throughout England and Scotland boarded up windows to avoid the tax.[5] In 1848, Mr. Byers, the president of the Carpenters' Society in London, reported to Parliament that nearly every house on Compton Street in Soho had employed him to reduce the number of windows (House of Commons 1848). In many houses, bricks took the place of previously existing windows. Moreover, newly constructed dwellings economized in drastic ways on the number of windows. In at least one apartment building in Edinburgh, the entire second floor (containing bedrooms) had no windows at all. Of course, there are some instances in which residents by design had numerous windows as a means of displaying their wealth.

The most serious adverse effect of the window tax was on human health. A series of studies by physicians and others found that the unsanitary conditions resulting from the lack of proper ventilation and fresh air encouraged the propagation of numerous diseases such as dysentery, gangrene, and typhus. In one instance in 1781, a typhus epidemic killed many citizens in Carlisle. Dr. John Heysham traced the origins of the outbreak to a house inhabited by six poor families (Guthrie 1867, p. 409), and described the dwelling in this way:

> In order to reduce the window tax, every window that even poverty could dispense with was built up, and all source of ventilation were thus removed. The smell in this house was overpowering, and offensive to an unbearable extent. There is no evidence that the fever was imported into this house, but it was propagated from it to other parts of town, and 52 of the inhabitants were killed.[6]

A series of petitions to Parliament resulted in the designation of commissioners and committees to study the problems of the window tax in the first half of the 19th century. In 1846, medical officers petitioned Parliament for the abolition of the window tax, pronouncing it to be "most injurious to the health, welfare, property, and industry of the poor, and of the community at large" (House of Commons 1850). Indeed, when Parliament acknowledged the serious damage to public health

on behavior). However, unlike Smith, Marshall speaks approvingly of the tax as a measure of ability-to-pay, arguing that the number of windows provides a reasonable index of "the scale and style of household expenditure in general" (p. 802). He contends: "If the part of the tax assessed on houses were removed, and the deficit made up by taxes assessed on the furniture and indoor servants, the true incidence of the taxes would be nearly the same as now" (p. 802). Marshall, along with Adam Smith in the passage quoted earlier in the text, fails to address the quite striking effects of such taxes on efforts to avoid their payment.

[5] There are many references to the window tax in English literature. In the 1748 novel *Tom Jones,* for example, one of Henry Fielding's characters exclaims (p. 380): "Why now there is above forty Shillings for Window-lights, and yet we have stopped up all we could; we have almost blinded the house I am sure . . ."

[6] One reviewer of this paper suggests that this opinion by a 19th century physician needs to be taken "with a grain of salt." This may be true, but as we note, there was widespread recognition of the injurious health effects of the window tax.

resulting from the blocking of windows, this ultimately resulted in the repeal of the tax in 1851.

Conceptual Framework

A tax system creates a "notch" if a small change in behavior leads to a discrete change in both average and marginal tax rates. As we noted above, the window tax incorporated notches throughout much of its history. Consider, for example, the tax schedule over the 1747–1757 period. As we showed above, a person who owned a home with 9 or fewer windows paid no tax. But his neighbor whose home had 10 windows would pay a tax of 6 pence *for each window.* Consequently, for the neighbor, the marginal tax rate for the 10th window was 60 pence (which is equal to 5 shillings) while the average tax rate for the 10 windows was 6 pence.

Notches are uncommon and have received relatively little attention in the literature on taxation (for an excellent overview of notches, see Slemrod 2010). "Kinks" are far more common. A tax system creates a kink if a small change in behavior leads to a discrete change in the marginal tax rate but just a very small change in the average rate. The United States federal individual income tax, for example, has several kinks. Earning an additional dollar could move a taxpayer into the next higher tax bracket, thus raising the marginal tax rate with (almost) no effect on the average tax rate. For an empirical study of bunching at kink points under the US income tax, see Saez (2010).

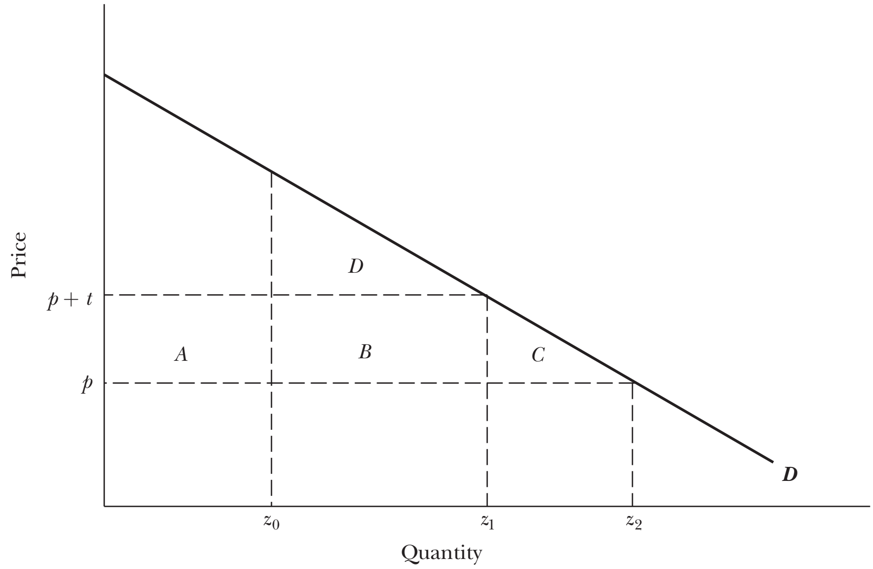
Public finance economists often argue against notches on the grounds that they lead to large deadweight losses: that is, a tax schedule with notches provides strong incentives for taxpayers to distort behavior and locate at a notch.[7] We explore this argument as we develop a conceptual framework to think about the window tax.

Consider a simple window tax that includes just one notch. Consumers pay no tax if they own z_0 or fewer windows but pay a tax of t pence per window if they own more than z_0 windows. In looking at Figure 1, there will be three cases to consider. Case I consists of consumers who would own fewer than z_0 windows in the absence of the tax. Case I consumers continue to own the same number of windows after the window tax is put in place. Thus, Case I consumers pay no tax and suffer no deadweight loss.

Case II consumers purchased more than z_0 windows before the tax and continue to purchase more than z_0 windows after the tax is imposed (though fewer windows than they did initially unless demand is perfectly inelastic). Figure 1 shows

[7] Blinder and Rosen (1985), however, argue that in some important cases, tax and subsidy plans with notches should at least be considered as serious contenders when public policy seeks to encourage or discourage some activity.

Figure 1
Demand for Windows and a Window Tax with a Single Notch



Notes: Consider a simple window tax that includes just one notch. Consumers pay no tax if they own z_0 or fewer windows but pay a tax of t pence per window if they own more than z_0 windows. Case I includes consumers who would own fewer than z_0 windows in the absence of the tax. Case I consumers pay no tax and suffer no deadweight loss. Case II consumers purchased more than z_0 windows before the tax and continue to purchase more than z_0 windows after the tax is imposed. They pay a total tax bill of $A + B$, suffer a welfare loss of $A + B + C$, and thus incur a deadweight loss of C. Case III includes consumers who would buy more than z_0 windows if there were no tax, but exactly z_0 once the tax is imposed. These consumers pay zero tax and suffer a welfare loss of $D + B + C$. Aside from Case I consumers, the decision on whether to pay the tax turns on the relative sizes of area D and area A.

the impact of the window tax on Case II consumers.[8] This consumer purchases z_2 windows at the market price p but z_1 windows once the tax is imposed. The notch is irrelevant for Case II consumers. For them, the window tax is equivalent to a standard excise tax of t pence per window. They pay a total tax bill of $A + B$, suffer a welfare loss of $A + B + C$, and thus incur a deadweight loss of C.[9]

Case III includes consumers who would buy more than z_0 windows if there were no tax, but exactly z_0 once the tax is imposed. These consumers pay zero tax. The tax has, however, distorted their decisions and so they suffer a deadweight loss.

[8] Formally, welfare losses should be calculated from the compensated (Hicksian) demand curve rather than the ordinary (Marshallian) demand curve. In practice, this distinction rarely turns out to be very important.

[9] In general, deadweight loss depends on both supply and demand. There is an implicit assumption throughout this paper that the supply curve for windows is perfectly elastic.

A Case III consumer owns $z_2 - z_0$ fewer windows as a result of the tax. Before the tax, the consumer realized consumer surplus of $D + B + C$ from those windows (the difference between willingness to pay and price), and so a Case III consumer suffers a welfare loss of $D + B + C$.

Which consumers fall into Case II, those who choose to pay the tax, and which into Case III, those who avoid the tax by restricting their consumption of windows? Case II consumers suffer a loss of $A + B + C$; Case III consumers suffer a loss of $D + B + C$. Consumers will choose the option that minimizes their loss from the tax. And so we come to the following rule: Consumers will choose to pay the tax (Case II) if $A + B + C < D + B + C$. They will avoid the tax (Case III) if $A + B + C > D + B + C$. They will be indifferent if $A + B + C = D + B + C$.

Subtracting $B + C$ from both sides shows that the key here is the relative magnitudes of areas A and D in Figure 1. The intuition behind this result is as follows. A consumer could choose to pay the tax and therefore purchase an additional $z_1 - z_0$ windows. The benefit from paying the tax is the difference between willingness to pay for windows and the price of a window (including the tax) integrated over $z_1 - z_0$ windows, area D. But in order to be able to purchase these $z_1 - z_0$ windows, the consumer must pay the tax on the first z_0 windows, area A. So the decision on whether or not to pay the tax turns on whether the benefit from purchasing additional windows (area D) is greater than, less than, or equal to the cost (area A).

This analysis suggests how to test the hypothesis that the window tax distorted people's decisions. If the window tax distorted decisions, then we should find "too many" people at the notches.[10] We present such a test in the next section of the paper.

How the Window Tax Distorted Decisions

To explore the quantitative impact of the window tax on actual behavior, we assembled a dataset from local tax records in 18th and 19th century Britain that indicates the number of windows per household over the period 1747 to 1830. We describe the dataset, and how we went about pulling it together, in the online Data Appendix to the paper available with this paper at http://e-jep.org.

We focus initially on the observations in our dataset from 1747 to 1757. As we discussed above, the window tax was unchanged over this period and included three notches. A homeowner in this period paid no tax if the house had fewer than 10 windows; a tax of 6 pence per window if the house had 10–14 windows; a tax of 9 pence per window if the home had 15–19 windows; or a tax of 1 shilling per

[10] More specifically, the test we outline here is a test of a sufficient condition that the tax distorted decisions. It is possible that if the notches were set so low that everyone purchased more windows than the number of untaxed windows, no one falls in Case III. The tax, in this example, would still distort decisions since the tax would have the same impact as a standard excise tax.

window if the home had 20 or more windows. (In addition, each homeowner paid a fixed duty of 2 shillings. This house tax was separated from the window tax in 1747.) Thus the marginal and average tax rate jumped sharply when a consumer installed the 10th, 15th, or 20th window.

We have tax data on 496 homes during this period. Most of the observations in our dataset are from Ludlow, a market town in Shropshire. Ludlow is close to the border with Wales. It had a population of roughly 4,000 people at the start of the nineteenth century; its current population is 10,500. We have data for two or more years for roughly 60 percent of the houses in our sample and for just a single year for the remaining 40 percent. We have treated our data as a cross section. In those cases where the number of windows changed over the 11-year period, we used the last observation available. We want to capture the effect of the tax, and using the last observation for each home gives us the greatest opportunity to observe a homeowner's response to the window tax.

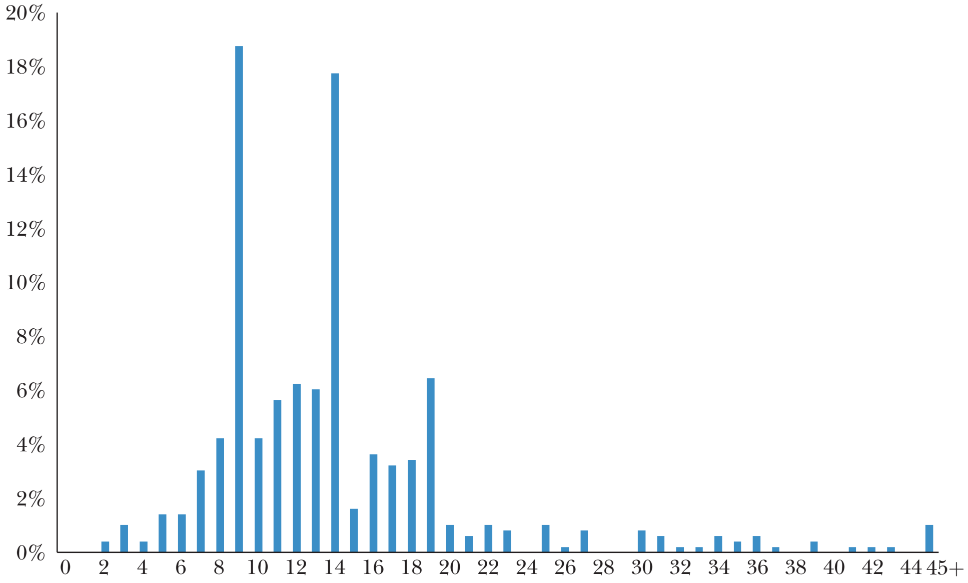
The period from 1747–1757 is a particularly useful sample for our purposes. As Glantz (2008) explains, the administration of the window tax proved to be an ongoing, difficult problem. It was common for homeowners to camouflage or board up windows until the tax collector was gone. Homeowners and local surveyors often avoided the window tax by taking advantage of loopholes and ambiguities in the tax code. The tax was imposed on every window in inhabited houses, while all industrial or retail buildings and homes of low-income families were exempt. Homeowners frequently attempted to disguise regular living quarters by storing a few sacks of grain in a room. Bribery and corruption among tax assessors was common.

As a result, tax collections were often much lower than expected. Parliament revised the window tax in 1747 to deal with these problems, and included heavy fines for attempts to evade the tax. As part of the 1747 act, "The practice of blocking up windows in order to evade assessment and subsequently reopening them, was prohibited under a penalty of 20s for every window reopened without due notice given to the tax surveyor" (Glantz 2008, pp. 8–9). These penalties were steep: a fine of 20 shillings is 20–40 times as large as the tax on windows itself. The 1747 revisions also included a number of provisions that improved the administration of the tax.

The 1747 act apparently was able to reduce tax evasion significantly. Data from the 1747–1757 period are therefore likely to yield a reasonable estimate of the actual number of windows. Data from earlier periods are more likely to reflect often successful efforts to evade the tax and therefore understate the actual number of windows.[11]

[11] In fact, some studies in other contexts have interpreted a large data value at a key cutoff as evidence of corruption. Stigler (1986, cited in Duggan and Levitt 2002), for example, showed that the height distribution among French males based on measurements taken at conscription was normally distributed except for a shortage of men measuring 1.57–1.597 meters (roughly 5 feet 2 inches to 5 feet 3 inches) and an excess number of men below 1.57 meters. Not coincidentally, the minimum height for conscription into the Imperial army was 1.57 meters.

Figure 2
Distribution of Number of Windows, 1747–1757 Sample



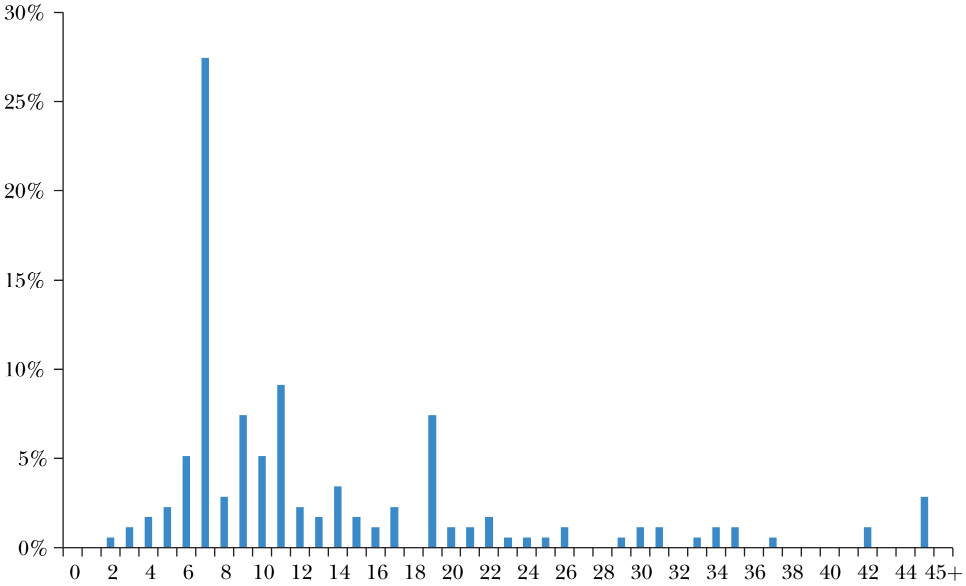
If the window tax distorted behavior, then we should expect to see "too many" homes with 9, 14, or 19 windows. This in fact is exactly what we find. Figure 2 presents a histogram showing the number of windows for homes in our sample. The pattern here is clear. There are sharp spikes in the number of homes at all three notches.[12] At the first notch, 18.8 percent of the homes have 9 windows, while 4.2 percent have 8 windows and 4.2 percent have 10 windows; at the second notch, 17.7 percent have 14 windows, while 6.0 percent have 13 windows and 1.6 percent have 15 windows; and at the third notch, 6.5 percent have 19 windows, while 3.4 percent have 18 windows and 1.0 percent have 20 windows.

Recall that the 1761 revisions to the window tax established a tax rate of 1 shilling per window on houses with 8 or 9 windows; from 1747 until 1760, only houses with 10 or more windows were subject to the tax. This change suggests a second test of the hypothesis that the window tax distorted people's decisions. We should expect to find "too many" houses with 7 windows beginning in 1761 but not in periods before 1761.

We collected a sample of 170 houses from the period 1761–1765 (there were significant changes to the tax rate in 1766). The houses in this second sample are from Wiltshire and Hampshire in southwest England. Figure 3 shows the distribution

[12] We present some straightforward statistical tests of the results in this section in the online Appendix available with this paper at http://e-jep.org.

Figure 3
Distribution of Number of Windows, 1761–65 Sample



of the number of windows for the homes in our 1761–1765 sample. We find a very large spike at 7 windows. In this sample, 27.4 percent of the houses have 7 windows but just 5.1 percent have six and just 2.9 percent have 8. In sharp contrast, just 3.0 percent of the houses in our 1747–1757 sample had 7 windows.

We also find concentrations in our 1761–65 sample at 11 windows (9.1 percent) and 19 windows (7.4 percent). This is consistent with 1761–65 tax policy; there were notches at both 11 and 19 windows during this period. In summary, the evidence from both samples is consistent with the hypothesis that property owners' decisions were distorted by the window tax. Our finding is in keeping with the observations of the prominent British historian M. Dorothy George (1926, p. 77), who noted: "When the duty was increased in 1710 it became a universal practice to stop up lights. How increasingly general the practice became may be gathered from the fact that in 1766 when the tax was extended to houses with 7 windows and upwards, the number of houses in England and Wales having exactly 7 windows was reduced by nearly two-thirds."

How Large Was the Deadweight Loss from the Window Tax?

We use a simulation model to develop a rough estimate of the deadweight loss from the window tax. We certainly would not claim that our simple model is able to capture all elements of tax policy in mid-18th century England. We would, however, argue that the model offers a sensible estimate of the order of magnitude of the efficiency cost of the tax.

We summarize the basic structure of the model here and present a more detailed discussion in the online Appendix. There are 1,000 consumers in the simulation. The price elasticity of demand is the same for all of the consumers but the height of their demand curves varies to reflect differences in incomes, tastes, and other determinants of the demand. The simulation first solves for the demand for windows in the absence of the tax. Each of the 1,000 consumers calculates consumer surplus (willingness to pay minus expenditure) if they were to buy 0, 1, 2, . . . 60 windows and chooses the number of windows that maximizes their consumer surplus. We then re-run the model under a tax policy that is similar to the 1747–1757 window tax. Consumers in our model who own 9 or fewer windows pay no tax; those who own 10–14 windows pay a tax of 6 pence per window; and those who own 15 or more windows pay a tax of 9 pence per window.[13] Each consumer in the model re-optimizes given this tax policy. The model captures each consumer's demand for windows with and without the tax; consumer surplus with and without the tax; and taxes paid.

We searched for values of the important parameters of the model that yield results that correspond most closely to our 1747–1757 data. Our estimated price elasticity of demand for windows is .149 (and so, for example, a 10 percent increase in the price of windows would reduce the demand for windows by 1.49 percent). We do not have any evidence against which we can evaluate this estimate. This estimate may seem low, but it is important to note that the demand for windows may be slow to adjust to a change in tax policy since the stock of new houses is small compared to the stock of existing houses (though as we argued above, many homeowners responded to the tax by blocking up existing windows).

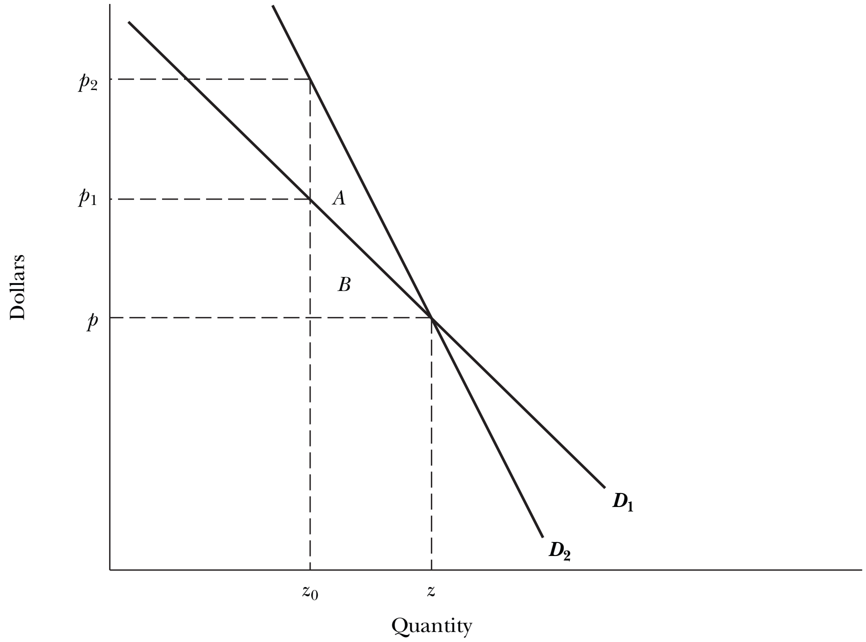
The magnitude of the price elasticity has some interesting implications for our estimate of the deadweight loss from the window tax.[14] Recall our earlier discussion of consumers who do not locate at a notch (Case II) and those who do (Case III). We argued that the cost of locating at a notch is the difference between willingness to pay for windows and the price of a window (excluding the tax) integrated over the windows a consumer foregoes by choosing the notch. When demand is inelastic and so the demand curve is steep, willingness to pay rises quickly as the quantity of windows falls. Therefore, if demand is inelastic, it is costly to choose to locate at a notch and so we should expect to find fewer consumers at a notch.

Figure 4 is helpful in seeing this result when demand is linear. Suppose consumers pay no tax for the first z_0 windows and a tax of t pence per window on all windows above z_0. If demand is relatively elastic (and so the relevant demand curve

[13] The tax policy we looked at in the simulation did not include the third notch that existed under the 1747–1757 tax (consumers with homes with 20 or more windows paid 1 shilling per window). Only 11 percent of the homes in our sample have 20 or more windows.

[14] We thank David Autor for his very helpful suggestions for this section of the paper.

Figure 4
Deadweight Loss from a Quantity Restriction



is D_1), choosing z_0 windows will lead to a deadweight loss of B. But if demand is relatively inelastic (and so the relevant demand curve is D_2), choosing z_0 windows leads to much larger deadweight loss of $A + B$.[15]

Figure 4 also suggests that the losses for consumers who do choose a notch—what we have called Case III—are large when demand is inelastic. For those consumers, the window tax is a quantity distortion. As Oates (1997) explains, the welfare loss from a policy that distorts quantity directly is large when demand is inelastic and small when demand is elastic; in the limit, the loss from a constraint on quantity is zero when demand is perfectly elastic. The intuition here is that where demand is less responsive, a consumer's valuation of marginal units rises quickly as we move away from the optimum.

The losses for consumers who do not locate at a notch—what we have called Case II—is straightforward. For this group, the window tax is a standard excise tax. The deadweight loss from an excise tax is small when the elasticity of demand is small. In the limit, the deadweight loss will be zero if demand is perfectly inelastic (because in that case a consumer's decision will be unaffected by the tax).

[15] Neary and Roberts (1980) would call p_1 or p_2 the shadow price of a window. A shadow price in this context is the price of a window that would lead a consumer to purchase z_0 windows in the absence of the quantity restriction.

Figure 5
Deadweight Loss from a Tax

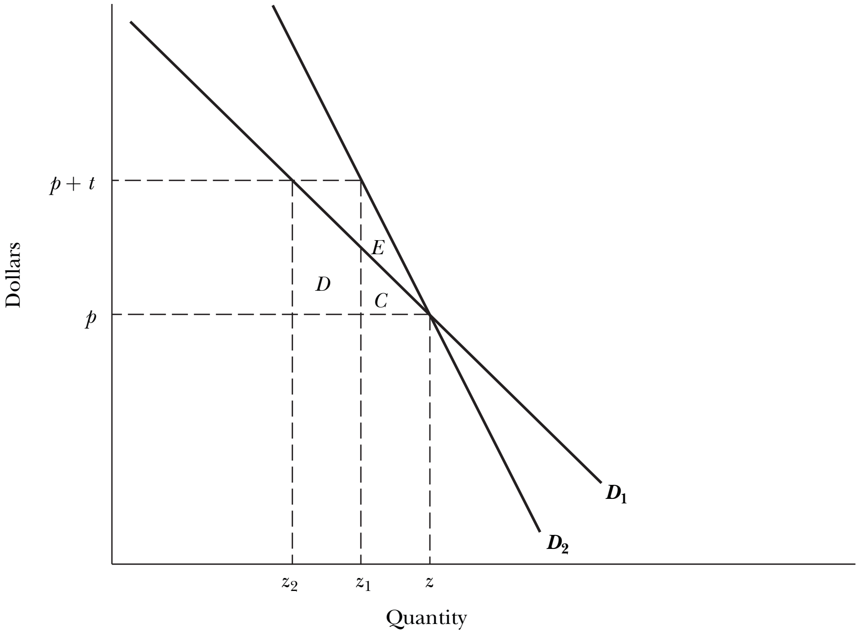


Figure 5 makes this point clear in the simple case where demand is linear. Demand curve D_1 is more elastic than demand curve D_2. A tax of t pence per window leads to a deadweight loss of $D + C$ if demand is elastic but just $E + C$ if demand is inelastic; to see this, note that $D + C$ equals $(t/2)(z - z_1)$ and $E + C$ equals $(t/2)(z - z_2)$.

This argument is similar in some ways to the Weitzman (1974) analysis of price and quantity instruments in environmental policy. In that paper he shows that a quantity instrument such as cap and trade is equivalent to a price instrument such as a Pigovian tax if the marginal abatement cost of pollution is known. Weitzman then considers the case where the marginal abatement cost is uncertain and, as a consequence, actual abatement costs turn out to be different from the regulator's estimate of abatement costs when either the regulatory price or quantity was chosen. He shows that the welfare effects of these two alternative instruments depend on the relative slopes of the marginal benefit curve from abatement and the marginal abatement cost curve. In particular, Weitzman argues that in the uncertainty case, a price instrument is more efficient than a quantity instrument when the marginal-benefit-of-abatement curve is relatively flat but that a quantity instrument is more efficient if the marginal benefit curve is relatively steep. As in our case of the window tax, the relationship between the slope of the demand curve (that is, the marginal benefit curve) and the magnitude of a distortion is different for price and quantity instruments.

Table 1
Simulation Results and 1747–1757 Data

	Simulation	*1747–1757 data*
Share of houses with 9 windows	21.9%	18.8%
Share of houses with 14 windows	13.3%	17.7%
Mean number of windows	14.1	14.1

Table 1 shows the actual and simulated values for the percentage of homes with exactly 9 windows, the percentage of homes with exactly 14 windows, and the mean number of windows. As that table shows, we were able to replicate the key features of our 1747–1757 data fairly closely. In the simulation, 21.9 percent of the homes had 9 windows, 13.3 percent had 14 windows, and the average number of windows is 14.1. In our 1747–1757, sample 18.8 percent had 9 windows, 17.7 percent had 14 windows, and the average home had 14.1 windows.

The window tax has a significant effect on the demand for windows in the simulation. None of the consumers in the simulation chose 10, 11, 15, or 16 windows when faced with the 1747–1757 window tax; it is never optimal to buy at a notch. The tax reduces average demand from an estimated 16.2 windows in the absence of the tax to an estimated 14.1 windows.

We focus initially on the consumers in the simulation who chose one of the notches. As we noted above, 21.9 percent of the households in the simulation chose 9 windows when faced with our version of the 1747–1757 tax schedule. This includes 5.5 percent of the sample that also chose 9 windows in the absence of the tax and whose choices were therefore not distorted. Thus 16.4 percent of the simulated households chose 9 windows under the tax, but more than 9 windows in the absence of the tax. All of the households that chose 14 windows when faced with the tax chose more than 14 windows in a world without the window taxes. And so in total, 29.7 percent of the households in the simulation chose one of the notches in direct response to the window tax.

How large is the distortion from the window tax? The estimated losses were very large for the households at one of the two notches. We find that for those consumers the deadweight loss equaled 62 percent of the taxes those consumers paid. That is to say, for every dollar collected the simulated version of the window tax imposed an additional burden of 62 cents on the households at the notches (over and above the direct burden of the tax paid). The excess burden, not surprisingly, is particularly large for households that chose 9 windows. Those consumers paid zero in window tax, and so for them the entire burden of the tax is excess burden.

We now turn to the entire sample of 1,000 simulated households. There are a number of alternative ways to think about the excess burden of a tax. We could focus on *total* excess burden as a fraction of *total* tax. In our simulation, the deadweight loss from the window tax is 13.4 percent of tax revenues. Alternatively, we might focus on the marginal excess burden of the window tax, which is a common measure of the distortionary effect of a tax. It is defined as the *marginal* excess burden from a *marginal* increase in tax revenue. We have calculated the marginal excess burden of the window tax by increasing the tax rates by 10 percent in the model and then calculating the resulting change in deadweight loss divided by the change in tax. We find a marginal excess burden of .23—raising an additional $1 of tax revenue through the window tax would generate an additional $0.23 of excess burden.

Concluding Remarks

The window tax provides a clear illustration of the deadweight loss from taxation. The discussion of deadweight loss can sometimes become a tangled debate over the measurement of Harberger triangles, partial versus general equilibrium estimates, and so on. Here is a clear case in which we mean what we say when we talk about excess burden. The window tax led many people to live in very dark houses and in environments that had significant, pernicious effects on their health.

The window tax is thus a quite striking example of a tax that led to radical tax-avoiding behavior with high associated levels of excess burden. This raises a further, intriguing question that goes beyond the scope of this paper but is worthy of mention here. If the window tax was a bad tax that generated such adverse effects and intense criticism, why did it persist over such a lengthy period? In fact, the rates were raised, in some instances quite dramatically.

The answer to this question requires a broader consideration of the political and fiscal issues of the times. But these were years of intense fiscal pressures, involving at various junctures massive military expenditures. The monarch and Parliament resorted in several instances not just to increases in the land and window taxes, but to a range of new taxes on various commodities and the introduction of an income tax (Dowell 1884, vol. 2). Thus, continued use of the window tax was, in part at least, a response to a setting of extreme budgetary tightness in which the government perceived little room for reduction in any tax rates. Perhaps the lesson here is that when governments need to raise significant revenue, even a very bad tax can survive for a very long time.

■ *We thank the Lincoln Institute of Land Policy for their generous support for this project. We gratefully acknowledge Andrew Glantz's very valuable contributions to our understanding of the history of the window tax and Katharine Hamilton's excellent work as our research assistant. We also thank the editors of this journal, Joan Youngman, Semida Munteanu, John Wallis, and Andrew Reschovsky for their very helpful comments, and Professor Michael Galgano of James Madison University who suggested very early on the likely availability of the relevant local historical data in the United Kingdom.*

References

Beckett, J. V. 1985. "Land Tax or Excise: The Levying of Taxation in Seventeenth- and Eighteenth-Century England." *English Historical Review* 100(395): 285–308.

Binney, J. E. D. 1958. *British Public Finance and Administration, 1774–92.* Oxford: Clarendon Press.

Blinder, Alan S., and Harvey S. Rosen. 1985. "Notches." *American Economic Review* 75(4): 736–47.

Dickens, Charles. 1850. *Household Words*, vol. 1. London: Bradbury and Evans.

Douglas, Roy. 1999. *Taxation in Britain Since 1660.* London: MacMillan.

Dowell, Stephen. 1884. *A History of Taxation and Taxes in England from the Earliest Times to the Present Day*, vols. 2 and 3. London: Frank Cass & Co. (Reprinted in 1965 in Reprints of Economic Classics series; New York: Kelley).

Duggan, Mark, and Steven D. Levitt. 2002. "Winning Isn't Everything: Corruption in Sumo Wrestling." *American Economic Review* 92(5): 1594–1605.

Fielding, Henry. 1749 [1975]. *The History of Tom Jones, A Foundling.* (Page numbers from the 1975 reprint, Wesleyan University Press.)

George, M. Dorothy. 1926. *London Life in the XVIIIth Century.* New York: Alfred A. Knopf.

Glantz, Andrew E. 2008. "A Tax on Light and Air: Impact of the Window Duty on Tax Administration and Architecture, 1696–1851." *Penn History Review* 15(2): Article 3.

Guthrie, Thomas. 1867. "How to Get Rid of an Enemy." *The Sunday Magazine for 1867,* p. 409. London: Strahan & Co.

House of Commons. 1819. "Motion for the Repeal of the Window Tax in Ireland," House of Commons Debates, May 5, 1819. In *Hansard House of Commons Debates,* vol. 40, cc 126–48.

House of Commons. 1848. "Lowest Classes Under Assessment," House of Commons Debates, February 24, 1848. In *Hansard House of Commons Debates*, vol. 96, cc 1259–97.

House of Commons. 1850. "Window Tax," House of Commons Debates, April 9, 1850. In *Hansard House of Commons Debates*, vol. 110, cc 68–99.

Kennedy, William. 1913. *English Taxation, 1640–1799.* London: G. Bell and Sons, Ltd.

Marshall, Alfred. 1890 [1948] *Principles of Economics*, Eighth Edition. (Page numbers from reprint, Macmillan, New York.)

Mateer, Dirk, and Lee Coppock. 2014. *Principles of Microeconomics.* New York: W. W. Norton & Company.

Neary, J. Peter, and Kevin S. W. Roberts. 1980. "The Theory of Household Behavior under Rationing." *European Economic Review* 13(1): 25–42.

Oates, Wallace E. 1997. "On the Welfare Gains from Fiscal Decentralization." *Journal of Public Finance and Public Choice* 2(3): 83–92.

Rosen, Harvey S., and Ted Gayer. 2010. *Public Finance*, Ninth Edition. McGraw-Hill Higher Education.

Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* 2(3): 189–212.

Sinclair, John. 1803. *The History of the Public Revenue of the British Empire.* London: Strahan and Preston.

Slemrod, Joel. 2010. "Buenas Notches: Lines and Notches in Tax System Design." http://webuser.bus.umich.edu/jslemrod/pdf/Buenas%20Notches%20090210.pdf.

Smith, Adam. 1776 [1937]. *The Wealth of Nations.* New York: Random House.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge, MA: Belknap Press of Harvard University Press.

Stiglitz, Joseph E. 1988. *Economics of the Public Sector,* Second Edition. New York: W. W. Norton.

Ward, W. R. 1952. "The Administration of the Window and Assessed Taxes, 1696–1798." *English Historical Review* 67(265): 522–42.

Weitzman, Martin L. 1974. "Prices *vs.* Quantities." *Review of Economic Studies* 41(4): 477–91.

Matthew Gentzkow, Winner of the 2014 Clark Medal

Andrei Shleifer

The 2014 John Bates Clark Medal of the American Economic Association was awarded to Matthew Gentzkow of the University of Chicago Booth School of Business. The citation recognized Matt's "fundamental contributions to our understanding of the economic forces driving the creation of media products, the changing nature and role of media in the digital environment, and the effect of media on education and civic engagement." In addition to his work on the media, Matt has made a number of significant contributions to empirical industrial organization more broadly, as well as to applied economic theory. In this essay, I highlight some of these contributions, which are listed on Table 1. I will be referring to these papers by their number on this list.

Matt earned both his AB in 1997, and, after a brief career in the theatre, his PhD in 2004 from Harvard, where he began to work on the media. At Harvard he also met Jesse Shapiro, his close friend and collaborator. I was one of Matt's (as well as Jesse's) thesis advisors. From Harvard, both Matt and Jesse moved to Chicago Booth School, where their research truly thrived and they contributed to a fantastic group of applied economists.

Background on Economics of the Media

After journalists played a prominent role in uncovering the Watergate conspiracies of the early 1970s, US newspapers for a time enjoyed an extraordinary

■ *Andrei Shleifer is Professor of Economics, Harvard University, Cambridge, Massachusetts. His email address is ashleifer@harvard.edu.*

http://dx.doi.org/10.1257/jep.29.1.181 doi=10.1257/jep.29.1.181

Table 1
Selected Papers by Matthew Gentzkow

1. “The Rise of the Fourth Estate: How Newspapers Became Informative and Why it Mattered,” (with Edward L. Glaeser and Claudia Goldin). 2006. Chap. 7 in *Corruption and Reform: Lessons from America’s Economic History,* edited by Edward L. Glaeser and Claudia Goldin. University of Chicago Press.
2. “Media Bias and Reputation,” (with Jesse M. Shapiro). 2006. *Journal of Political Economy* 114(2): 280–316.
3. “Television and Voter Turnout.” 2006. *Quarterly Journal of Economics* 121(3): 931–72.
4. “Valuing New Goods in a Model with Complementarity: Online Newspapers.” 2007. *American Economic Review* 97(3): 713–44.
5. “Preschool Television Viewing and Adolescent Test Scores: Historical Evidence from the Coleman Study,” (with Jesse M. Shapiro). 2008. *Quarterly Journal of Economics* 123(1): 279–323.
6. “What Drives Media Slant? Evidence from U.S. Daily Newspapers” (with Jesse M. Shapiro). 2010. *Econometrica* 78(1): 35–71.
7. “Bayesian Persuasion,” (with Emir Kamenica). 2011. *American Economic Review* 101(6): 2590–2615.
8. “Ideological Segregation Online and Offline,” (with Jesse M. Shapiro). 2011. *Quarterly Journal of Economics* 126(4): 1799–1839.
9. “The Effect of Newspaper Entry and Exit on Electoral Politics,” (with Jesse M. Shapiro and Michael Sinkinson). 2011. *American Economic Review* 101(7): 2980–3018.
10. “The Evolution of Brand Preferences: Evidence from Consumer Migration,” (with Bart Bronnenberg and Jean-Pierre Dubé). 2012. *American Economic Review* 102(6): 2472–2508.
11. “Competition and Ideological Diversity: Historical Evidence from US Newspapers,” (with Jesse M. Shapiro and Michael Sinkinson). 2014. *American Economic Review* 104(10): 3073–3114.
12. “Competition in Persuasion,” (with Emir Kamenica). 2011. NBER Working Paper 17436.
13. “Do Pharmacists Buy Bayer: Sophisticated Shoppers and the Brand Premium,” (with Bart Bronnenberg, Jean-Pierre H. Dubé, and Jesse M. Shapiro). 2013. Chicago Booth Research Paper No. 14-17. Available at SSRN: http://ssrn.com/abstract=2460893.

reputation for objectivity. Just as intended by the US Constitution, the narrative ran, the Fourth Estate found and reported the unvarnished truth about dishonest and corrupt politicians and brought it to the attention of voters. Newspapers, of course, used editorial pages to express opinions, but the news sections stuck to the facts. This reputation of the media was bolstered by the protections for freedom of the press in the US Constitution, various pieces of regulation, as well as Supreme Court rulings that made it close-to-impossible to win a lawsuit for libel against a newspaper (because such a lawsuit had to prove malice and reckless disregard for the truth, rather than just falsehood and negligence). While economists developed rather cynical views of politics (with public choice theory) and of regulation (with regulatory capture theory), they mostly bought into a “normative” model of the press. Even as economists accumulated theories and evidence on self-serving behavior of politicians and regulators, they left the study of the press—the profit-seeking, competitive press—largely to First Amendment scholars.

Matthew Gentzkow

Over the following decades, this image of the media began to change. One important development was the publication in the early 2000s of a series of fire-breathing exposés by right-wing journalists (Coulter 2003; Goldberg 2003) and left-wing journalists (Alterman 2003) accusing the media as a whole of extreme reporting slant. The right saw a left-wing slant; the left detected an equally pernicious right-wing slant. These books, as well as the growing prominence of television stations often accused of political partisanship—like CNN and Fox News—led some economists to become interested in the objectivity of the media. Several obvious questions stood out.

First, is media news reporting actually slanted? Is it the case that, editorial pages aside, media outlets report unbiased news, or alternatively, through commission or omission, do they deliberately bias their reporting?

Second, if reporting is biased, what is the reason? Is such bias driven by the supply-side, as when reporting reflects the prejudices of an outlet's owners or journalists? Indeed, the journalistic accounts of the media bias in the early 2000s took it for granted that the biases of owners and reporters drove the slant. Alternatively, is the slant driven by demand, as when news outlets cater to the preferences of their audience to maintain or increase their readership or viewership?

Third, what is the effect of media competition on accuracy and bias? Does competition increase the accuracy of reporting by individual outlets so even a consumer of only one source gets better information? Alternatively, does competition make it easier for the "whole truth" to come out from the perspective of a hypothetical reader who samples many media sources even when individual outlets are biased? In this regard, does a typical media consumer rely on one source of news or seek truth by sampling a variety of sources?

Finally, does media reporting actually matter for individual understanding and action? Does it affect knowledge? Does it influence participation in the political process? Does it influence how people vote? Do television and newspapers have the same effects, or different ones?

In a very short decade, economic research has obtained fairly clear answers to at least some of these questions. To a large extent, this is the consequence of Gentzkow's work, both on his own and with Jesse Shapiro. In the process, economists have arrived at a much deeper and more thorough understanding of the workings of the Fourth Estate, leaving First Amendment scholars behind.

Media Bias

A pair of theoretical papers published in the mid-2000s clarified the role of competition in shaping media bias when newspapers cater to the demand of their readers. In Mullainathan and Shleifer (2005), my coauthor and I consider the case of two profit-maximizing newspapers deciding where to locate on a segment of readers differentiated by their ideological preferences. In our model, by assumption, readers consciously trade off accuracy of a news source against a preference for information that confirms their beliefs. As a consequence, readers are willing to pay more for a newspaper whose slant reflects their own bias. In equilibrium, competition leads newspapers to cater to biased readers through slant.

The underlying logic can be understood in terms of the famous Hotelling (1929) model, which begins with an example of two producers facing a set of consumers evenly distributed along a segment, which Hotelling analogizes both to a geographic segment like Main Street in a town and also to an ideological segment like the political spectrum. Hotelling argues that if the consumers will give their business to whoever is nearest to them, then the two producers will have an incentive to cluster in the middle of the segment. If one producer moved either left or right, that producer would lose customers as the other producer would move in that same direction while just remaining on the longer side of the segment. In contrast, if the producers can charge more if they move closer to their customers, then instead of clustering, the producers have an incentive to choose separate locations. In Mullainathan and Shleifer, we show that in a competitive equilibrium with two newspapers, both will report biased news to readers who are willing to pay for slant, but with opposite ideological slants. In fact, adding additional newspapers would lead to segregation of readers across sources closest to their biases, and might lead to reduced accuracy of individual outlets. At the same time, a reader exposed to all sources will obtain more accurate information through averaging out the slants.

Unlike in Mullainathan and Shleifer (2005), where we simply assume a taste for confirming news even if coverage is inaccurate, Gentzkow and Shapiro [2] make the more subtle, and perhaps more plausible, assumption that readers rationally prefer sources consistent with their priors because they sincerely believe that those sources are more accurate. They report (p. 286) the results of a survey in which "nearly 30 percent of the respondents who described themselves as conservative indicated that they thought they could believe all or most of what the Fox Cable News Network says. In contrast, less than 15 percent of self-described liberals said that they could believe all or most of what the network reports." Gentzkow and Shapiro then build a model in which newspapers slant the news toward the priors of their readers to establish a reputation for quality. As in Mullainathan and Shleifer, this model generates slanted reporting in equilibrium. However, the model predicts that competition reduces such bias, because inaccurate reporting would damage newspaper reputation in the long run.

These theoretical models helped clarify some of the basic issues on newspaper competition. Still, they would have been remained mere theoretical possibilities had Gentzkow and Shapiro [6] not written a wonderful empirical paper in 2010 examining the sources of media bias. The challenge was to measure the political orientation of different newspaper outlets, which in modern days all see themselves as independent. At the casual level, Gentzkow and Shapiro recognized that the words newspapers use reflect their bias. For instance, on May 18, 2004, the liberal *Washington Post* headline read "After Decades of Courting and Waiting, Same-Sex Couples Line Up Early for a Marriage Made in Massachusetts." On the same day, the conservative *Washington Times* headline read "Homosexuals 'marry' in Massachusetts." As a more recent example, consider the November 30, 2013, headlines as the US government rushed to repair the HealthCare.gov website. According to the *Washington Post*, "HealthCare.gov will meet deadline for fixes, White House Officials say." According to *USA Today*, "Deadline's here: Is Healthcare.gov fixed? Sort of." According to the *Wall Street Journal*, "Health Site Is Improving But Likely to Miss Saturday Deadline." But how can one turn these kinds of anecdotes into data?

Building on the work of Groseclose and Mylo (2005), Gentzkow and Shapiro [6] found a solution. They developed a measure of media slant based on the proximity of an outlet's language to that of Republicans and Democrats in Congress, using a dataset of all the phrases in the Congressional Record in 2005 categorized by the party of the speaker of the phrase. In 2005, for example Democrats in Congress disproportionately referred to a "war in Iraq," while Republicans referred to a "war on terror." Gentzkow and Shapiro then collected data on the use of these highly diagnostic phrases in US daily newspapers and used these data to place news outlets on the ideological spectrum comparable to members of Congress.

In addition to this large methodological advance in how to measure partisan newspaper slant, the paper used detailed information on newspaper circulation and voting patterns across space to estimate a model of the demand for slant and to show that—consistent with the theory—consumers gravitate to like-minded sources, giving the newspapers an incentive to tailor their content to their readers. They also show that newspapers respond to that incentive and that variation in reader ideology explains a large portion of the variation in slant across US daily newspapers.

As important, Gentzkow and Shapiro [6] show that, after controlling for a newspaper's audience, the identity of its owner does not affect its slant.[1] Two newspapers with the same owner look no more similar in their slant than newspapers with different owners. Ownership regulation in the US and elsewhere is based on the premise a news outlet's owner determines how it spins the news. Gentzkow and Shapiro produced the first large-scale test of this hypothesis, which showed that, contrary to the conventional wisdom and regulatory stance, demand is much more influential in shaping content than supply as proxied by ownership.

US newspaper markets today mostly have a single major newspaper, so to look at the effects of competition between newspapers on ideology, Gentzkow, Shapiro, and Michael Sinkinson [11] turn to the past. At the turn of the 20th century, many US cities had multiple competing newspapers, and newspapers commonly expressed explicit partisan affiliations. Gentzkow, Shapiro, and Sinkinson assembled the US Newspaper Panel, a complete census of English-language daily newspapers in all presidential election years from 1872–2004. They also collected geographically disaggregated data on newspaper circulation in 1924, as well as income statements from a small sample of newspapers. Using these data, they estimate a quantitative model of newspaper competition in which news outlets compete for both readers and advertisers.

An important aspect of the Gentzkow, Shapiro, and Sinkinson [11] estimation strategy is to deal with complementarity and substitution between different outlets. They rely on and extend an earlier paper of Gentzkow's [4], which looked at competition between print and online newspapers, and in particular examined the question of whether print and online versions of the same newspaper are complements or substitutes. Gentzkow found that print and online newspapers are substitutes, and measured the magnitude of crowding out from the introduction of online versions. In the process, he developed a tractable framework for discrete-choice demand in which consumers buy bundles of products rather than single items.

In Gentzkow, Shapiro, and Sinkinson [11], the authors find that competition is a key driver of ideological diversity: newspapers differentiate ideologically as a strategy to soften competition for advertisers and for readers, in line with theoretical models. They also find that the market undersupplies diversity, in the sense that a policymaker concerned with consumer and producer welfare would want more markets in which readers can choose to read both a local Republican paper and a local Democratic paper. Interestingly, they consider two kinds of subsidies for newspapers: subsidies for newspaper distribution of the sort first created by the Postal Act of 1792, which continued to be important to newspapers at least up through the 1920s, and the indirect subsidies provided by the Newspaper Preservation Act

[1] Of course, the Gentzkow and Shapiro [6] paper is focused on private newspapers in the United States today. In other countries, newspapers and television stations are often either owned (Djankov, McLiesh, Nenova, and Shleifer 2003) or subverted (Besley and Prat 2006) by the government, so politicization of the press is a much greater issue. In the United States historically, newspapers were affiliated with political parties and toed the party line (Gentzkow, Glaeser, and Goldin [1]).

of 1970, which allow newspapers in a city to sign a joint operating agreement that often combines the business operations of the two papers while keeping the news operations separate. They find that such subsidies can improve the functioning of the market for news, delivering more value to the participants in the market as well as more diversity in the marketplace of ideas.

These empirical studies of newspaper competition answer quite conclusively the first two questions: news reporting is indeed slanted, and the principal reason for slant is catering to reader demand. Unbiased news reporting is a myth, not the reality, of US media.

The third question—do readers end up exposed only to biased news?—is harder to answer, since it requires knowing the extent to which readers are exposed to one or multiple sources of news. To address this question, Gentzkow and Shapiro [8] move away from newspapers and study the effect of the Internet on the ideological diversity of the American news diet. One might worry that the increase in choice among news suppliers as a result of the Internet would allow news consumers to self-segregate, reading only news that confirms their preconceptions. Gentzkow and Shapiro test this claim using data from a panel of Internet users for which they have a survey-based measure of political ideology and tracking data on online news consumption. They find that ideological segregation is surprisingly low online. The average conservative's news outlet on the Internet is about as conservative as usatoday.com; the average liberal's is as liberal as cnn.com. Strikingly, the Internet is less ideologically segregated than US residential geography: two people using the same news website are less likely to have an ideology in common than two people living in the same zip code.

Effects of the Media

Measuring media bias and understanding the interplay between industry competition and ideology in the media industry are important accomplishments. Of course, we also want to know whether the media, biased or otherwise, has any effect on politics. For example, does reading a newspaper or watching television make people more likely to vote? In addition, does the bias of the news sources actually affect how viewers vote? For obvious reasons, these questions are difficult to answer. Readers of newspapers might vote because they are stimulated by newspapers to participate. Or prospective voters might read newspapers because they seek information. Or some factor such as an interest in politics, either general or partisan, could drive both newspaper reading and voting. For example, Fox News might persuade people to vote Republican or, alternatively, Republican voters might choose to watch Fox News.

One solution to these identification problems is to focus on (preferably exogenously determined) entry—or exit—by news organizations into local markets, and to examine its consequences on the amount and type of voting. Gentzkow has also been a pioneer in this line of research. In Gentzkow [3], he uses variation across markets in the timing of the introduction to television in the United States to identify its impact on voter turnout. He estimates a huge negative effect: the availability of television accounts for between one-quarter and one-half of the total decline in voter turnout since the 1950s. Matt argues that a principal reason for this is substitution in media consumption away from newspapers, which provide more political coverage and thus stimulate more interest in voting. In line with this conjecture, he shows "that the entry of television in a market coincided with sharp drops in consumption of newspapers and radio" as well as a decline "in political knowledge as measured by election surveys." Also "both the information and turnout effects were largest in off-year congressional elections, which receive extensive coverage in newspapers but little or no coverage on television" (p. 931).

Gentzkow, Shapiro, and Sinkinson [9] use their US Newspaper Panel to consider the effects of newspapers on voting. Specifically, they focus on entries and exits of US daily newspapers between 1869 and 2004 to estimate effects on voter turnout and voter partisanship. They find that newspapers have a large effect in raising voter turnout, especially in the period before the introduction of broadcast media. However, the political affiliation of entering newspapers does not affect the partisan composition of an area's vote. The latter result contrasts with another important finding, by DellaVigna and Kaplan (2007), that the entry of Fox News does sway some voters toward voting Republican. An interpretation consistent with these findings is that newspapers motivate but don't persuade, while television does the opposite.

Another follow-up study partially redeems television, although in a nonpolitical sphere. Gentzkow and Shapiro [5] "use heterogeneity in the timing of television's introduction to different local markets to identify the effect of preschool television exposure on standardized test scores during adolescence" (p. 279). Contrary to conventional wisdom, watching TV makes you smarter: "an additional year of preschool television exposure raises average adolescent test scores by about .02 standard deviations" (p. 294). Based on my own teenage experience, I am particularly sympathetic to their finding that these positive effects of television on test scores "are largest for youngsters from households where English is not the primary language" (p. 279).

Economics of Brands and Branding

Consumer brands raise fascinating issues for economics. Why are consumers attached to some brands that they then buy repeatedly? Why do they pay a premium for brands? Do brands represent superior products or are they just trusted by consumers who could buy equally good unbranded items?

Bronnenberg, Dubé, and Gentzkow [10] present some remarkable facts about brand loyalty by looking at consumers who move from one city to another. They show that movers continue to buy the brands they bought in their places of previous residence, even if their new city is dominated by another brand. The paper shows that brand preferences form endogenously based on where consumption started, are highly persistent, and explain 40 percent of geographic variation in market shares. Put differently, there are Coke cities and Pepsi cities, and people growing up in a Coke city would continue to drink Coke even if they move to a Pepsi city. Brand preferences are almost addictive.

Bronnenberg, Dubé, Gentzkow, and Shapiro [13] address a different question about brands: do brands reflect superior objective quality? They ask whether specialists, such as doctors or chefs, buy branded products or generic ones. They find that although even experts often buy branded products, experts are much more likely than nonexperts to buy generics and avoid brands. They interpret this finding as suggesting that branding is a mechanism for conveying quality information to uninformed buyers, information for which these buyers willingly pay. This quality information is already known to experts, who therefore do not need to pay for it.

Economics of Persuasion

Persuasion has been central to economics beginning at least with Stigler's (1961) work on advertising, which interpreted advertising as provision of information to potential buyers. Two decades later, Grossman and Hart (1980), Milgrom (1981), and Milgrom and Roberts (1986) proved a paradoxical result about persuasion. If the persuader has information that the audience does not have, and the persuader cannot lie, then the persuader will have to disclose truthfully all of the information that the persuader has, for failure to disclose any individual item would be interpreted as hiding the worst facts. Ugly truth is better than selective omission, since the latter means the truth is even uglier. The finding appears to suggest that, with rational parties, persuasion through selective disclosure of information does not work: the best one can do is tell the whole truth.[2]

Kamenica and Gentzkow [7] take a fresh look at this problem, reframe it, and obtain some quite unexpected results. Rather than focusing on the persuader with superior information, they ask a different question: supposing the persuader and the audience begin with the same information, can the persuader design a test, which the audience will see the results of, that would actually further the goals of the persuader? In concrete terms, can a prosecutor look for evidence, with the judge knowing exactly what type of evidence the prosecutor is looking for, how the prosecutor is looking for it, and what the prosecutor finds, that will make the judge more likely to convict? Can an advertiser design a "taste test," with the potential customers knowing exactly what the advertiser is doing, that would increase demand?

[2] A Spring 2008 symposium on "The Economics of Persuasion" in this journal features Paul Milgrom's summary of his work in this area, Matthew Gentzkow and Jesse Shapiro's overview of the research on accuracy in media, and Peter Leeson's cross-country study of the relationship between media freedom, political knowledge, and participation.

At first glance the answer might seem to be "no." Indeed, there is a precise sense in which persuasion is difficult in such settings: a Bayesian audience cannot expect to be surprised, so its expected posterior is always equal to the prior. Thus, a persuader interested in changing the audience's average posterior is out of luck.

What Kamenica and Gentzkow [12] show is that the "no-surprise-on-average" property of the Bayes' Rule summarizes all the restrictions. With the right choice of tests, the persuader can in principle achieve any distribution of posterior beliefs on the part of the audience subject to the "no-surprise-on-average" constraint. This insight yields a beautiful geometric representation of the persuader's problem. It allows Kamenica and Gentzkow to show that if the persuader has a payoff that is nonlinear in the audience's belief, then persuasion is possible in the sense that the persuader can design a test that furthers the goals of the persuader. It also allows for a precise characterization of the optimal testing regime for a variety of interesting problems.

To take a specific example, suppose the murderer's blood is left at the crime scene. We know the defendant has blood type A. Suppose that the judge's and the prosecutor's prior belief that the defendant is guilty of murder is .3; their information is completely symmetric. Suppose the judge convicts if the posterior probability of guilt is above .5, so absent an investigation the judge would always acquit. If, instead, the prosecutor publicly conducts a fully informative investigation that perfectly reveals guilt, he can increase the prior odds of conviction from 0 to 30 percent, that is, convicting all the guilty and acquitting all the innocent: because the judge's action is nonlinear in beliefs, the prosecutor can benefit from providing full information despite the "no-surprise-on-average" constraint.

Perhaps more surprisingly, Kamenica and Gentzkow [7] show that the prosecutor can do even better by performing a less-informative investigation. To illustrate, the prosecutor proposes to the judge to test the type of blood at the crime scene. If the defendant is indeed guilty, the crime-scene blood is always type A: $\Pr(A|\text{Guilty}) = 1$. If the defendant is innocent, the crime-scene blood is of type A 42 percent of the time, given blood type frequencies in the US population: $\Pr(A|\text{Innocent}) = .42$. With this test, the posterior probability of guilt is just above .5 whenever the test indicates type A blood, so the judge convicts whenever the test comes back type A. More precisely, by Bayes' Rule,

$$\begin{aligned}\Pr(\text{Guilty}|A) &= [\Pr(A|\text{Guilty}) * \Pr(\text{Guilty})]\backslash[\Pr(A|\text{Guilty}) * \Pr(\text{Guilty}) \\ &\qquad + \Pr(A|\text{Innocent}) * \Pr(\text{Innocent})] \\ &= [1 * .3]/[1 * .3 + .42 * .7] > .5.\end{aligned}$$

With a prior of .3 of guilt, this test, if conducted and reported truthfully, yields a prior probability of conviction of $.3 * 1 + .7 * .42 = .594$. With symmetric beliefs, and the judge and the prosecutor both knowing exactly what is going on, the prosecutor can raise the odds of conviction all the way to 60 percent despite the parties knowing that only 30 percent of the defendants are guilty. In this very precise way, persuasion is effective.

In follow-up work, Gentzkow and Kamenica [12] extend this analysis to the case of multiple persuaders, who choose what information to gather and communicate to a receiver who can take actions that affect their welfare. They show that competition among persuaders necessarily increases the amount of information being revealed. This result connects with the earlier finding of Gentzkow and Shapiro [2] that competition among news outlets necessarily increases accuracy.

Summary

Ten years ago, we knew almost nothing about how newspapers actually report news. There were questions, but no answers—just media hype. Today, we actually have answers to many of the questions that were raised initially. We know that media reporting is systematically slanted, that slant is largely driven by demand, and that competition allows more of the viewpoints to get out. We also know that media influence their audiences for sure in getting them to participate in politics and sometimes in how they vote as well. At the same time, we have many new questions about the media: How exactly do they persuade? How do readers decide how many sources to attend to? How will the rise of new digital platforms and revenue models affect media content and political discourse? That media economics is now a full-fledged field is significantly a consequence of the contributions made by Matthew Gentzkow.

After rereading Matt's papers, and reading some for the first time, I am struck by his openness to different ways of doing economics. He has an uncanny ability to rely on different approaches, depending on what the problem he is considering calls for. Sometimes he uses quasi-experimental evidence to identify the effects he is interested in; other times he estimates full structural models. Some papers deal with small data sets; others rely on frontier big data techniques. Several of the papers contain practical econometric advances that have become useful to subsequent researchers. Sometimes Matt uses the simplest models that only summarize the verbal ideas; other papers, such as the work on persuasion, contain significant contributions to economic theory. Much of his work is extremely neoclassical, but some is behavioral as well. Some papers deal with abstract conceptual issues; others are solidly grounded in practical concerns, including regulatory ones.

This range is admirable not just for its own sake. My sense is that when areas of economics conclude that there is only one correct way of analyzing a problem, they stagnate. Our discipline is not far enough along to settle down in this way. Openness to new ways of doing things is essential for making progress. I would go further and conjecture that such openness is the hallmark of 21st century economics. The fact that Matthew Gentzkow along with his remarkable collaborators and several other recent winners of the John Bates Clark Medal embrace such openness is both a testimony to their talents and very good news for our field.

■ *I am grateful to Emir Kamenica, Jesse Shapiro, and Josh Schwartzstein for extremely helpful comments.*

References

Alterman, Eric. 2003. "*What Liberal Media? The Truth About Bias and the News.*" Basic Books.

Besley, Timothy, and Andrea Prat. 2006. "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability." *American Economic Review* 96(3): 720–36.

Coulter, Ann. 2003. "S*lander: Liberal Lies About the American Right.*" New York: Three Rivers Press.

DellaVigna, Stefano, and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics* 122(3): 1187–1234.

Djankov, Simeon, Caralee McLiesh, Tatiana Nenova, and Andrei Shleifer. 2003. "Who Owns the Media?" *Journal of Law and Economics* 46(2): 341–81.

Goldberg, Bernard. 2003. "*Bias: A CBS Insider Exposes How the Media Distort the News.*" New York: Perennial.

Groseclose, Tim, and Jeffrey Milyo. 2005. "A Measure of Media Bias." *Quarterly Journal of Economics* 120(4): 1191–1237.

Grossman, Sanford J., and Oliver D. Hart. 1980. "Disclosure Laws and Takeover Bids." *Journal of Finance* 35(2): 323–34.

Hotelling, Harold. 1929. "Stability in Competition." *Economic Journal* 39(153): 41–57.

Milgrom, Paul. 1981. "Good News and Bad News: Representation Theorems and Applications."*Bell Journal of Economics* 12(2): 380–91.

Milgrom, Paul, and John Roberts. 1986. "Relying on the Information of Interested Parties." *RAND Journal of Economics* 17(1): 18–32.

Mullainathan, Sendhil, and Andrei Shleifer. 2005. "The Market for News." *American Economic Review* 95(4): 1031–53.

Stigler, George J. 1961. "The Economics of Information." *Journal of Political Economy* 69(3): 213–25.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 193–206

Retrospectives
The Marginal Cost Controversy

Brett M. Frischmann and Christiaan Hogendorn

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please write to Joseph Persky of the University of Illinois at Chicago at jpersky@uic.edu.

Introduction

From 1938 to 1950, there was a spirited debate about whether decreasing-average-cost industries should set prices at marginal cost, with attendant subsidies if necessary. In 1938, Harold Hotelling published a forceful and far-reaching proposal for marginal cost pricing entitled "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates." After several years and many pages of discussion, Ronald Coase gave a name and a clear formulation to the debate in his 1946 article "The Marginal Cost Controversy." We will tell much of the story of this controversy by comparing the frameworks of Hotelling and Coase, while also bringing in other contributors and offering some thoughts about contemporary relevance. The arguments marshaled by Coase (and his contemporaries) not only succeeded in this particular debate, as we shall see,

■ *Brett M. Frischmann is Professor and Director of the Intellectual Property and Information Law program, Cardozo Law School, Yeshiva University, New York City, New York. Christiaan Hogendorn is Associate Professor of Economics, Wesleyan University, Middletown, Connecticut. Their email addresses are frischma@yu.edu and chogendorn@wesleyan.edu.*

http://dx.doi.org/10.1257/jep.29.1.193 doi=10.1257/jep.29.1.193

but more generally served as part of the foundation for various fields of modern economics, particularly institutional, regulatory, and public choice economics as well as law and economics. Yet the underlying issues are quite difficult to resolve, and the strengths and weaknesses of the arguments for marginal cost pricing can turn on specific elements of the industry.

The Case for Marginal Cost Pricing

The origins of the marginal cost controversy can be traced back to a discussion in Book V, Chapter XII of Alfred Marshall's (1890) *Principles of Economics* (as explained in Ruggles 1949a). Marshall pointed out that in what he called an "increasing returns" industry—in which marginal costs of production were falling at the quantity relevant to market demand—having the government paying a "bounty" (a subsidy) to the producers could benefit consumers. The bounty would shift out the supply curve, which with declining marginal costs would bring down the price, thus expanding consumer surplus. Marshall wrote: "[A] bounty on such a commodity causes so great a fall in its price to the consumer, that the consequent increase of consumers' surplus may exceed the total payments made by the State to the producers; and certainly will do so in case the law of increasing return acts at all sharply."

Indeed, Marshall offered a further illustration that J. H. Clapham (1922) famously challenged as an "empty economics box." Marshall showed a graphical example in which a revenue-neutral combination of a tax on an increasing-marginal-cost industry combined with a bounty for a decreasing-marginal-cost industry could raise consumer surplus. A. C. Pigou, Allyn Young, J. H. Clapham, Knight, and others argued over this result until Clapham and then Knight pointed out that with static technology, the decreasing-marginal-cost industry was not creating a positive externality, but merely benefitting from internal economies of scale in related firms. Thus, there was no market failure and Marshall's tax-subsidy scheme would not actually increase welfare. By the 1930s, this controversy was dying down, and economists turned away from industry-level decreasing costs to internal decreasing costs at the firm level. In particular, a common focus was the example of large firms with high fixed costs and low marginal costs, like railroads. In such industries, the average cost of production was declining over a substantial range of output, with the low marginal costs falling below the average costs over that same range of output. As a result, a price set equal to marginal cost would not cover the average cost of production, and would cause a firm to sustain losses.

Hotelling (1938) brought many of these arguments together.[1] He appealed to the basic economic intuition that efficiency requires marginal cost pricing because

[1] Ruggles (1949a) describes how Hotelling's (1938) essay built on a number of slightly earlier works. For example, Dickinson (1933) retained the old idea about taxing increasing-marginal-cost industries and introduced the criterion of pricing at marginal cost for decreasing-marginal-cost industries. Abba

it fulfills the efficiency condition that social welfare will be greatest when marginal benefits (as captured in the price consumers are willing to pay) are equated to marginal costs. Hotelling argued that "the optimum of the general welfare corresponds to the sale of everything at marginal cost" and that general government revenues should "be applied to cover the fixed costs of electric power plants, waterworks, railroad, and other industries in which the fixed costs are large, so as to reduce to the level of marginal cost the prices charged for the services and products of these industries" (p. 242).

What about the concern that setting price equal to marginal cost would not allow these industries to cover their fixed costs and thus would force them to sustain losses? In the jargon of the day, a project that generated the revenue to cover its fixed costs was said to be "self-liquidating." Hotelling (1938) wrote: "The notion that public projects should be 'self-liquidating,' on which President Hoover based his inadequate program for combatting the oncoming depression, while attractive to the wealthier tax-payers, is not consistent with the nation's getting the maximum of satisfactions for its expenditure" (p. 260).

Thus, Hotelling (1938) suggested that government subsidization of the fixed cost component would enable marginal cost pricing for industries with high fixed and low marginal costs. Where would the government obtain the needed revenue for the necessary subsidies? Hotelling refers back to Jules Dupuit's (1844) result that excise taxes cause what we now call deadweight loss, indeed that the deadweight loss is proportionate to the *square* of the tax rate. Hotelling proved the result mathematically using a consumer's utility maximization problem and commented, "[I]f a person must pay a certain sum of money in taxes, his satisfaction will be greater if the levy is made directly on him as a fixed amount than if it is made through a system of excise taxes which he can to some extent avoid by rearranging his production and consumption" (p. 252).

Having established the superiority of lump-sum taxes, Hotelling (1938) pointed out that excise taxes have the same undesirable traits as markups above marginal cost. Whether the reason for a markup above marginal costs is an excise tax, the need to recover fixed costs, or pure market power, the result is the same: deadweight loss and a lower level of social welfare. Thus, government revenue should come from lump-sum taxes, which could be used both to abolish excise taxes and to provide any needed subsidies to enable marginal-cost pricing. Hotelling mentioned five potential sources of these lump-sum, nondistortionary taxes: land, on-peak railway trips, advertising (because he claimed total time available for viewing advertising is fixed), inheritance, and income. All five suggestions were controversial, although thinking of an income tax in lump-sum terms probably proved the most contentious at the time. However, as Nancy Ruggles (1949b)

Lerner (1934) then stated that maximum social welfare occurs when a monopolist sets price equal to marginal cost, and R. F. Kahn (1935) extended this to say that assuming equal marginal utility of income, maximum welfare implies subsidizing decreasing-average-cost firms to enable marginal cost pricing.

noted, Hotelling's suggestions were actually fairly conservative and careful by the standards of the time, when other advocates of marginal cost pricing were often taking more radical positions involving nationalization of firms and extensive wealth redistribution.

When Prices Don't Include Opportunity Costs

Coase (1946) challenged Hotelling (1938) and others taking a similar position on the benefits of having marginal cost pricing with government subsidies to cover fixed costs in an industry with declining average costs. Coase (1946) expressed surprise that "despite the importance of its practical implications, its paradoxical nature, and the fact that there are many economists who consider it fallacious, it [the Hotelling proposal] has so far received little written criticism" (pp. 169–70). Coase acknowledged that when price is not equal to marginal cost, there is an efficiency loss. However, he wrote that Hotelling's proposal would "bring about a maldistribution of the factors of production, a maldistribution of income and probably a loss similar to that which the scheme was designed to avoid" (p. 180).

Coase (1946) began with an "isolation of the problem" that helped to clarify the issues. He set up the discussion in this way (p. 170):

> The central problem relates to a divergence between average and marginal costs. But, in any actual case, two other problems usually arise. First, some of the costs are common to numbers of consumers and any consideration of the view that total costs ought to be borne by consumers raises the question of whether there is any rational method by which these common costs can be allocated between consumers. Secondly, many of the so-called fixed costs are in fact outlays which were made in the past for factors, the return to which in the present is a quasi-rent, and a consideration of what the return to such factors ought to be (in order to discover what total costs are) raises additional problems of great intricacy.

As a conceptual simplifier, Coase introduced a scenario where there is a central market from which a carrier can make a radial journey, bringing any quantity at one fixed cost. Thus, the marginal cost of carrying an additional quantity is zero, even though there is a positive fixed cost. Coase argued that in the radial market, a Hotelling-style approach to marginal cost pricing would lead to a price of zero for carriage, with the government subsidizing the fixed costs. In this example, Coase noted three problems with the Hotelling scheme.

First, a violation of pricing principles seems to arise. If prices are set at marginal cost while fixed costs are subsidized, then both consumers and producers will not be taking fixed costs into account in their decision-making. Consider a situation in which a producer might either make something on-site, or instead purchase that input in the center of the radial market and have it transported at a price of zero. Coase (1946, pp. 172–73) argued that the governing rule for optimal pricing, specifically that "the amount paid for a product should be equal to its cost," must include the opportunity cost of factors. As Coase (1947) wrote in a follow-up essay: "If certain factors of production can be obtained free in one use (because they do not enter into marginal cost) but have to be paid for in another use (because they enter into marginal cost), consumers may choose to employ these factors in the use in which they are free even though they would in fact prefer to employ them in some other way" (p. 150). In the radial market scenario, if carriage is provided freely, a wedge will arise between prices and total costs that distorts consumer decision making and would lead to a "maldistribution of the factors of production between different uses" (p. 174).

A second concern is that government payment for carriage is only justified if consumers would have been willing to pay the full cost. But how can anyone know if that condition holds if carriage is not priced? This brings up the issue of how to know what fixed costs the government should be subsidizing when everything is priced at marginal cost—a question we take up in the next section.

The third concern Coase (1946) pointed out in the radial scenario is that if everyone pays the same price for carriage, there will be redistribution of wealth from high-density (usually urban) users of infrastructure who would otherwise have benefitted from low fixed costs per user toward lower density users who in some other arrangement would have paid much higher fixed costs per user.

Coase (1946) argued in favor of multi-part pricing because it allows the consumer to be charged separately for products (units purchased) and for carriage of products from the central market—that is, in two parts. Coase was not especially clear about how the multipart price might be determined. While the radial market avoids certain kinds of complexity, he admits that it also assumes away the key difficulty of *common* costs of carriage. Coase (1947) advocated cost-based differential pricing rather than value-based price discrimination and criticized Nordin (1947) for misreading his work to conclude that he favored value-based price discrimination. But he never fully addressed the problem of allocating common costs that have no obvious relationship to a particular customer.

Coase (1946) also argued that even average cost pricing can have advantages over marginal cost pricing. On the negative side, it does cause deadweight loss relative to marginal cost, and it does not provide a full test of overall willingness to pay. But avoiding the need for taxes may offset these inefficiencies. Also there is no need for government estimates of demand. Coase argued that these tradeoffs suggest that there should be no presumption in favor of marginal cost pricing. Overall, Coase argued that some form of multipart pricing could allow for average cost recovery while minimizing distortions from pricing above marginal cost. This multi-part pricing alternative (typically involving a fixed fee and a per-unit fee) would retain some degree of market-based demand signals, allow consumers to choose rationally how to spend their money, and generate better information to guide infrastructure investments on the supply-side.

Investment Tests

Subsidized marginal cost pricing—which can be understood as subsidies directed at fixed costs—eliminates or at least truncates signals about demand for infrastructure, significantly reducing the information available for investment decisions about how much infrastructure to build, where to build it, when to add capacity, and so on. With marginal cost pricing and a government subsidy that covers fixed costs, users do not receive a signal of the total cost of the good, and producers do not receive information about willingness to pay for the full cost of new goods or improvements. Under this condition, how will society make decisions about planning and investment in infrastructure markets, including yet-to-be-discovered and discovered-but-yet-to-be-built infrastructure markets?

Hotelling (1938) alluded to this investment test critique. He began by saying: "Determination whether to build the bridge by calculation merely of the revenue $\Sigma p_i q_i$ obtainable from tolls is always too conservative a criterion" (p. 248). He ended with a brief discussion of how to decide whether something new should be built. For the case of the railroads, Hotelling stressed that they are already built and so this question is moot: "Whether it was wise for the government to subsidize and its backers to construct the Union Pacific Railroad after the Civil War is an interesting historical question which would make a good subject for a dissertation, but it would be better, if necessary, to leave it unsolved than to ruin the country the Union Pacific was designed to serve by charging enormous freight rates and claiming that their sum constitutes a measure of the value to the country of the investment" (p. 268). As for future construction, Hotelling waved at the problem by saying that willingness to pay becomes a problem for "economists, statisticians, and engineers, and perhaps for a certain amount of large-scale experimentation . . ." (p. 269).

That rather blithe attitude toward the investment test touched off a great deal of criticism. Wilson (1945) objected that there was no general test of whether investment was justified—thus making future plans difficult. Coase (1946) argued that in his radial market example, government carriage is only justified if consumers would have been willing to pay the full cost. But how can anyone know if consumers would have been willing to pay that price if carriage is not priced? Relying on the government to subsidize the fixed cost component in decreasing-cost industries raises significant concerns about institutional competence. How will the government know when and who and how much to subsidize? How will the government determine what costs constitute the fixed costs?

From a static efficiency perspective focused exclusively on an already existing public utility, the Hotelling (1938) argument for marginal cost pricing has some clear merits. But when considering a dynamic perspective over time and a range of potential products, Coase (1946) argued that the case for government to pay for fixed costs seems weaker. He expressed skepticism that government would be able to ascertain individual preferences about the appropriate level of fixed costs to subsidize. Indeed, Coase suggested that if government could do so as needed to effectuate Hotelling's proposal, then there would be little need for markets and the pricing system in general.

In this and other writings, Coase was skeptical of rote confidence in government institutions, and he challenged economists to evaluate critically claims that relied on the expertise, competence, benevolence, and public-mindedness of government officials.

The General Equilibrium Critique and Redistribution Issues

Relying on the government to subsidize fixed costs also raises concerns about distortions caused by government taxation. Coase (1946) maintained that the impact of increasing income or other taxes to raise revenue would be substantial and could not be ignored. Similarly, while Abba Lerner (1944) favored Hotelling-style marginal cost pricing, he insisted that labor market effects of income taxation must be taken into account. This critique concerns general economy-wide distortions originating from taxation.

A Hotelling-style reliance on government taxation and spending also raises concerns about interpersonal comparisons and redistribution of wealth from the general population to public utility consumers. As Pegrum (1944) noted, consumers who benefit under the scheme were not necessarily identical to the taxpayers who paid for the fixed costs.

Hotelling (1938) addressed this question by pointing out that the public policy issue is not whether a single decreasing-cost firm should have its fixed costs subsidized, but whether a large number of such firms would be treated in this way. In any given case, some consumers would benefit more than others. But Hotelling argued: "A rough randomness in distribution would be ample to ensure such a distribution of benefits that most persons in every part of the country would be better off by reason of the programme as a whole." Coase (1946) responded by arguing: "But this argument stands or falls by the assumption that there will be no significant redistribution as between consumers of different kinds of products. There is no reason to assume that this will be so. . . . It would be possible to appraise the character of the redistribution only after a detailed factual enquiry. . . . I see no reason to suppose that there would not be some redistribution, possibly very considerable, as a result of this policy if it were generally applied."

Ruggles (1949b) both summarized various aspects of the earlier debate and made an important contribution that distills two potential general equilibrium problems with the Hotelling proposal: 1) taxes to fund subsidies would violate some marginal conditions, typically on the labor–leisure tradeoff; and 2) income would be redistributed, hence forcing interpersonal comparisons of utility. She argued that these objections are not fatal to a Hotelling-style proposal if and only if taxes fall on the consumer surplus of the actual consumers of the product. If a marginal cost pricing scheme passes this "Ruggles test," then the general equilibrium concerns are allayed. But otherwise, policymakers would (implicitly) have to revert to an assumption of equal marginal utility of income across consumers and a willingness to choose higher social welfare even if it involves some redistribution to justify a policy of marginal cost pricing.

The general equilibrium critiques of Lerner (1944), Coase (1946), and Ruggles (1949b) seem to have ended discussion (at least in the United States) of employing marginal cost pricing throughout the entire economy. But this still left the "public utilities," the specific industries in which the likelihood of natural monopoly and other market-failure concerns had prompted government ownership or regulation. Hotelling had specifically based his 1938 paper on the similarity between the problems of taxation and those of railway rate making; he also mentions electricity and water; and all applied examples at the end of Hotelling's article concern bridges and railroads.

Vickrey's Take

William Vickrey (1948) set out to revive the marginal cost pricing proposal, at least for public utilities. In essence, Vickrey argued that the critiques of marginal cost pricing actually point up the general difficulties and opportunities inherent in any decreasing-average-cost industry.

For example, Vickrey (1948) acknowledged that Coase's (1946) argument convincingly supports multipart pricing when all costs can be attributed to individual users. However, Vickrey argues that common costs, where it is very difficult to attribute costs to users, are the most important case, and Coase's proposal for multipart pricing in his radial market scenario sidesteps the difficulty of apportioning common costs. In many examples, like the case of the Tennessee Valley Authority, it is possible to estimate future marginal costs conditional on certain facilities being built. On the other hand, estimating future average cost requires difficult cost allocation between flood control, navigation, and electricity generation. Vickrey (1948) argues that "in a decreasing-cost industry, 'marginal cost' is a definite concept, though it may be difficult to measure, while 'average cost' for a specific type or date of output may be completely arbitrary, though accountants may be able to compute it with great accuracy in accordance with their more or less arbitrary rules" (p. 232).

As for the investment test critique, Vickrey (1948) agreed this was a significant problem but argued that it is always a problem for *any* decreasing-average-cost industry regardless of the pricing system. Multipart pricing often requires the same problematic levels of information as marginal cost pricing does; in contrast to Coase's radial system, Vickrey points out the complexities of apportioning costs that arise in a circuit delivery service. Vickrey writes: "[I]t is necessary to distinguish carefully between multipart schedules designed to extract a larger fraction of value of the service from the consumer and multipart schedules designed to reflect more accurately the marginal cost of a service having several parameters." Likewise, "it should not be thought that marginal-cost pricing would necessarily be uniform. . . . The issue is not one of relative complexity of rate schedules, but of the purpose that these complexities are designed to serve" (p. 220). Vickrey points out that requiring self-liquidation to avoid mistaken investment introduces a "substantial bias" against undertaking projects.

Vickrey's most ambitious proposal to implement his ideas was his 1952 proposal for a restructuring of the New York City subway fare structure (Vickrey 1952, 1955). He discussed marginal cost pricing, which implied low or zero fares in the direction against the rush hour, low fares for off-peak and short-haul traffic in the outer boroughs, and high prices for peak trips on the most congested routes. The total revenue would still not cover total costs, so general government revenue would need to make up the difference. Vickrey sought to justify this use of general government revenues based on both the high consumer surplus for users of the service and general benefits to the city from expanded commerce, lower congestion, and environmental considerations.

Aftermath

The marginal cost controversy was never fully settled. Both Vickrey (1970 [1994]) and Coase (1970) were still working on it decades later. In practice, the answer to the controversy seems to be a theoretical admission that marginal cost pricing would be socially efficient in certain industries with declining average costs and low marginal costs, coupled with a pragmatic argument that subsidizing fixed costs in these industries is politically difficult and so regulatory policy for declining-cost public utilities will often need to set prices above marginal cost.

Reflecting the theoretical admission, marginal cost remained the baseline for efficient pricing in textbook discussions. For example, in the 1988 edition of Alfred Kahn's prominent book on *The Economics of Regulation*, Chapter 3 is devoted to marginal cost pricing and begins with strong endorsement of the concept: "The central policy prescription of microeconomics is the equation of price and marginal cost. If economic theory is to have any relevance to public utility pricing, that is the point at which the inquiry must begin" (p. 65).

However, the thrust of pricing policy toward the regulated industries—like electricity, natural gas, telephone, airlines, railroads—in the third quarter of the twentieth century did not involve much in the way of subsidy from general government revenue. Thus the emphasis was on optimal pricing, subject to the self-liquidation constraint that each firm must cover its own total costs. Coase (1970) noted that even in post–World War II Britain, with its many nationalized industries, the government did not implement marginal cost pricing with attendant subsidies. Duffy (2004) summarized the dominant approach:

> Modern regulatory policy generally accepts that a declining average cost industry—that is, a so-called "natural monopoly"—will not have its fixed costs subsidized from general government revenues and that therefore the industry must be allowed to price above marginal cost so that it can cover its fixed costs. The rejection of the Hotelling thesis is so complete that reputable economics encompasses the very opposite of Hotelling's view—"that, generally, prices which deviate in a systematic manner from marginal costs will be required

> for an optimal allocation of resources, even in the absence of externalities." Indeed, in the parlance of public utility regulation, the very phrase "marginal cost pricing" now refers not to Hotelling's proposed marginal cost pricing and subsidy scheme, but rather to a pricing system akin to the "multi-part" pricing system that Coase advocated as the more efficient alternative to Hotelling's proposal. In short, modern public utility theorists generally do not recommend using pervasive public subsidies to chase the Holy Grail of global marginal cost pricing.

Of course, one result of the self-liquidation constraint that utilities must cover their own costs was that pricing had to deviate from marginal cost. This in turn raised questions about the best way to cover fixed costs, whether through some form of average cost pricing or another approach. Baumol and Bradford (1970) applied principles going back to Ramsey (1927) that for greatest efficiency, prices should deviate from marginal cost in inverse proportion to demand elasticity. Ramsey pricing was one solution to the common-cost allocation problem that Coase had struggled with, but Ramsey prices are value-based—that is, they are based on demand for different products—and their use can imply considerable redistribution of income. They are, however, subject to a profit or break-even constraint, which limits the conversion of consumer surplus to producer surplus (Frischmann 2012, p. 16).

Another result was that policymakers tended to deal with externalities and social goals on an industry-by-industry basis. This meant that regulators had to decide how to allocate common costs; for example, telephone regulators had to allocate capital costs when setting long distance and local rates. Sometimes such pricing policies involved significant cross-subsidies between different services. In the telephone case, long distance subsidized local service for many years in order to promote widespread adoption of the telephone. In a competitive marketplace, an overpriced service would have been subject to competitive entry, so cross-subsidies often had to be accompanied by entry restrictions (Faulhaber 1975).

Interestingly, parallel developments in infrastructure outside the traditional regulated industries sometimes did involve something closer to Hotelling's (1938) approach, though generally without the "marginal cost pricing" nomenclature. In the United States, the most important instance was the (mostly) toll-free Interstate Highway System. In general, the marginal cost of an additional vehicle to the highway system is near-zero, with marginal costs associated with degradation being related to the number of miles driven and gasoline consumed; thus, highways were funded primarily through taxes on gasoline with some contribution from other sources of government revenue (Button 2010). Begun in 1956, the government created wide, fast, and relatively safe highways connecting communities across the nation. This critical infrastructure investment contributed to the growth of the economy by interconnecting markets, lowering the cost of transporting goods and people, and improving connectivity between communities close and distant.

Technological Complementarity, Productive Users, and Spillovers

Hotelling (1938), Coase (1946), and the other participants in the marginal cost controversy implicitly assumed that users of infrastructure were passive consumers operating in an unchanging, complete market without externalities. One exception arises when Hotelling (1938) and Vickrey (1948) mention the Tennessee Valley Authority. Given the large-scale positive externalities that Tennessee Valley electrification and flood control would generate, Hotelling argued that it would be better to sell the electricity at its marginal cost and make up the difference with revenues derived from other parts of the country.

Yet the Tennessee Valley Authority example is hardly exceptional. Many infrastructures generate positive externalities. Modern economics recognizes that infrastructure resources are nonrival or partially nonrival inputs into a wide variety of private, public, and social goods. Consumers of such infrastructures are not passive; instead, their resulting productive activities generate the spillovers (Frischmann 2012). The potential social gains here are substantial. The consumer surplus from introducing a new good, which Romer (1994) suggests should be named the "Dupuit triangle," is much larger than the deadweight loss triangles caused by slight departures from optimal pricing for existing goods. Similarly, Lipsey, Carlaw, and Bekar (2005) emphasized that the majority of spillovers caused by general purpose technologies are not marginal positive externalities, but instead involve what they term "technological complementarities."

The issues raised in the marginal cost controversy remain relevant but become more complicated where the assumption of passive consumers operating in an unchanging, complete market is relaxed. For example, marginal cost pricing issues are prominent in the modern arguments over government subsidization of fixed costs of certain information and communication technology infrastructures as well as government regulations that involve nondiscrimination rules for the Internet (the so-called "net neutrality" debate) (Hogendorn 2012).

In the last three decades, the Internet has grown to become an essential national infrastructure. It has reshaped commerce and increased entrepreneurship, as well as affected political discourse, the production and consumption of media, social network formation, and community building (Frischmann 2012). Decisions made in coming years regarding radio spectrum allocation, government investment in broadband and wireless infrastructure, and regulation of privately owned Internet infrastructure will have a direct, significant impact on its future.

A modern-day Hotelling might point out that when a general purpose infrastructure technology supports a number of complementary innovations, the concern with deadweight losses associated with pricing above marginal cost becomes even more pressing. In this situation, it is doubtful whether a multipart pricing scheme would reveal demand not only for the infrastructure in a narrow sense, but also for the eventual complements that would come into being as a result of that infrastructure. This modern-day Hotelling would doubtless point out that even though government subsidies of such technology impose costs on the general public by taxation, it also may confer widespread general welfare benefits as well. Indeed, the spillover effects from the information and communications technology often involve benefits that flow to those who are not using that technology directly themselves.

Naturally, a modern-day Coase would respond to these arguments by raising various issues and concerns: how alternative multipart pricing strategies might work as a way of paying for such investments; the deadweight losses that would be imposed by taxes to pay for any subsidies; the danger that if fixed costs are subsidized, prices will not reflect opportunity costs and will lead to distortions; and of course the risk that a politically elected government and its regulatory agencies may lack the competence to identify and manage such investments. Thus, modern arguments over public policy in industries with declining average costs are in many ways a reprise and updating of the original marginal cost controversy.

■ *We are grateful to Gerald Faulhaber, Michael Lovell, and Ingo Vogelsang for very thoughtful comments. All errors remain ours.*

References

Baumol, William J., and Bradford, David F. 1970. "Optimal Departures From Marginal Cost Pricing." *American Economic Review* 60(3): 265–83.

Button, Kenneth. 2010. *Transport Economics,* 3rd ed. Edward Elgar.

Clapham, J. H. 1922. "Of Empty Economic Boxes." *Economic Journal* 32(127): 305–314.

Coase, Ronald H. 1946. "The Marginal Cost Controversy." *Economica* 13(51): 169–82.

Coase, Ronald H. 1947. "The Marginal Cost Controversy: Some Further Comments." *Economica* 14(54): 150–53.

Coase, Ronald H. 1970. "The Theory of Public Utility Pricing and Its Application." *Bell Journal of Economics and Management Science* 1(1): 113–28.

Dickinson, H. D. 1933. "Price Formation in a Socialist Community." *Economic Journal* 43(170): 237–50.

Duffy, John F. 2004. "The Marginal Cost Controversy in Intellectual Property." *University of Chicago Law Review* 71(1): 37–56.

Dupuit, Arsène Jules Étienne Juvénal. 1844. "De la mesure de l'utilité des travaux publics." *Annales des Ponts et Chaussées* Second series, 8 (November). Translated by R. H. Barback as "On the Measurement of the Utility of Public Works," *International Economic Papers,* 1952, vol. 2, pp. 83–110.

Faulhaber, Gerald R. 1975. "Cross-Subsidization: Pricing in Public Enterprises." *American Economic Review* 65(5): 966–977.

Frischmann, Brett M. 2012. *Infrastructure: The Social Value of Shared Resources.* Oxford University Press.

Hogendorn, Christiaan. 2012. "Spillovers and Network Neutrality." Chap. 8 in *Regulation and the Performance of Communication and Information Networks,* edited by Gerald R. Faulhaber, Gary Madden, and Jeffrey Petchey. Edward Elgar.

Hotelling, Harold. 1938. "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates." *Econometrica* 6(3): 242–69.

Kahn, Alfred E. 1988. *The Economics of Regulation: Principles and Institutions.* MIT Press.

Kahn, R. F. 1935. "Some Notes on Ideal Output." *Economic Journal* 45(177): 1-35.

Lerner, A. P. 1934. "Economic Theory and Socialist Economy." *Economic Theory and Socialist Economy* 2(1): 51-61.

Lerner, Abba P. 1944. *The Economics of Control.* New York: Macmillan.

Lipsey, Richard G., Kenneth I. Carlaw and Clifford T. Bekar. 2005. *Economic Transformations: General Purpose Technologies and Long-term Economic Growth.* Oxford University Press.

Marshall, Alfred. 1890. *Principles of Economics.* Available online at the Archive for the History of Economic Thought: http://www.marxists.org/reference/subject/economics/marshall/.

Nordin, J. A. 1947. "The Marginal Cost Controversy: A Reply." *Economica* 14(54): 134–49.

Pegrum, D. F. 1944. "Incremental Cost Pricing: A Comment." *Journal of Land & Public Utility Economics* 20(1): 58–60.

Ramsey, Frank P. 1927. "A Contribution to the Theory of Taxation." *Economic Journal* 37(145): 47–61.

Romer, Paul M. 1994. "New Goods, Old Theory, and the Welfare Costs of Trade Restrictions." *Journal of Development Economics* 43(1): 5–38.

Ruggles, Nancy. 1949a. "The Welfare Basis of the Marginal Cost Pricing Principle." *Review of Economic Studies* 17(1): 29–46.

Ruggles, Nancy. 1949b. "Recent Developments in the Theory of Marginal Cost Pricing." *Review of Economic Studies* 17(2): 107–126.

Vickrey, William. 1948. "Some Objections to Marginal-Cost Pricing." *Journal of Political Economy* 56(3): 218–38.

Vickrey, William S. 1952. "The Revision of the Rapid Transit Fare Structure of the City of New York." Technical Monograph no. 3, Finance Project, Mayor's Committee for Management Survey of the City of New York.

Vickrey, William S. 1955. "A Proposal for Revising New York's Subway Fare Structure." *Journal of the Operations Research Society of America* 3(1): 38–68.

Vickrey, William. 1970 [1994]. "Airline Overbooking: Some Further Solutions." Chap. 13 in *Public Economics: Selected Papers by William Vickrey*, edited by Richard Arnott, Kenneth Arrow, Anthony Atkinson, and Jaques Drèze. Cambridge University Press.

Wilson, T. 1945. "Price and Outlay Policy of State Enterprise." *Economic Journal* 55(220): 454–61.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 207–214

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.

Potpourri

The *2015 World Development Report* from the World Bank, with a theme of "Mind, Society, and Behavior," offers a useful overview of the way in which these issues of "behavioral economics" affect the welfare of low-income people around the world, especially in the context of poverty, early childhood development, household finance, productivity, health, and climate change. Here's one example of many: "Fruit vendors in Chennai, India, provide a particularly vivid example. Each day, the vendors buy fruit on credit to sell during the day. They borrow about 1,000 rupees (the equivalent of $45 in purchasing parity) each morning at the rate of almost 5 percent per day and pay back the funds with interest at the end of the

■ *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.29.1.207 doi=10.1257/jep.29.1.207

day. By forgoing two cups of tea each day, they could save enough after 90 days to avoid having to borrow and would thus increase their incomes by 40 rupees a day, equivalent to about half a day's wages. But they do not do that. . . . Thinking as they always do (automatically) rather than deliberatively, the vendors fail to go through the exercise of adding up the small fees incurred over time to make the dollar costs salient enough to warrant consideration." The report includes evidence that development professionals are subject to these biases, too. "Dedicated, well-meaning professionals in the field of development—including government policy makers, agency officials, technical consultants, and frontline practitioners in the public, private, and nonprofit sectors—can fail to help, or even inadvertently harm, the very people they seek to assist if their choices are subtly and unconsciously influenced by their social environment, the mental models they have of the poor, and the limits of their cognitive bandwidth. They, too, rely on automatic thinking and fall into decision traps. Perhaps the most pressing concern is whether development professionals understand the circumstances in which the beneficiaries of their policies actually live and the beliefs and attitudes that shape their lives . . ." December 4, 2014. At http://www.worldbank.org/en/publication/wdr2015.

The OECD discusses "Challenges of International Co-operation in Competition Law Enforcement." "The number of cross-border cartels revealed in an average year has increased substantially since the early 1990s. According to the Private International Cartel (PIC) database, about 3 cross-border cartels were revealed via competition authority decisions or prosecutions in an average year between 1990 and 1994. In recent years, from 2007 to 2011, an average of about 16 cross-border cartels has been revealed per year. . . ." The spread of competition law enforcement around the world has been remarkable. At the end of the 1970s only nine jurisdictions had a competition law, and only six of them had a competition authority in place. . . . As of October 2013, about 127 jurisdictions had a competition law, of which 120 had a functioning competition authority. . . . The speed and breadth of the proliferation of competition laws and competition enforcers around the globe is the single most important development in the competition area over the last 20 years." 2014, at http://www.oecd.org/daf/competition/challenges-international-coop-competition-2014.htm.

The Financial Stability Board has published the *Global Shadow Banking Monitoring Report 2014*. "The 'shadow banking system' can broadly be described as 'credit intermediation involving entities and activities (fully or partially) outside the regular banking system' or non-bank credit intermediation in short. Such intermediation, appropriately conducted, provides a valuable alternative to bank funding that supports real economic activity. But experience from the crisis demonstrates the capacity for some non-bank entities and transactions to operate on a large scale in ways that create bank-like risks to financial stability . . . Like banks, a leveraged and maturity-transforming shadow banking system can be vulnerable to 'runs' and generate contagion risk, thereby amplifying systemic risk. Such activity, if unattended, can also heighten procyclicality by accelerating credit supply and asset price increases during surges in confidence, while making precipitate falls in asset prices and credit more likely by creating credit channels vulnerable to sudden loss of confidence. These effects were powerfully revealed in 2007–09 in the dislocation of asset-backed commercial paper (ABCP) markets, the failure of an originate-to-distribute model employing structured investment vehicles (SIVs) and conduits, 'runs' on MMFs [money market funds] and a sudden reappraisal of the terms on which securities lending and repos were conducted. But whereas banks are subject to a well-developed system of prudential regulation and other safeguards, the shadow banking system is typically subject to less stringent, or no, oversight arrangements. . . . [N]on-bank financial intermediation grew by $5 trillion in 2013 to reach $75 trillion. This provides a conservative proxy of the global shadow banking system, which can be further narrowed down. October 30, 2014. At http://www.financialstabilityboard.org/wp-content/uploads/r_141030.pdf.

In the 2014 Martin Feldstein lecture, Stanley Fischer asks "Financial Sector Reform: How Far Are We?" "The capital ratios of the 25 largest banks in the United States have risen by as much as 50 percent since the beginning of 2005 to the start of this year, depending on which regulatory ratio you look at. For example, the tier 1 common equity ratio has gone up from 7 percent to 11 percent for these institutions. . . . At the same time, the introduction of macroeconomic supervisory stress tests in the United States has added a forward-looking approach to assessing capital adequacy, as firms are required to hold a capital buffer sufficient to withstand a several-year period of severe economic and financial stress." "What about simply breaking up the largest financial institutions? Well, there is no 'simply' in this area. . . . Would a financial system that consisted of a large number of medium-sized and small firms be more stable and more efficient than one with a smaller number of very large firms? . . . That is not clear, for Lehman Brothers, although a large financial institution, was not one of the giants—except that it was connected with a very large number of other banks and financial institutions. Similarly, the savings and loan crisis of the 1980s and 1990s was not a TBTF [too big to fail] crisis but rather a failure involving many small firms that were behaving unwisely, and in some cases illegally. This case is consistent with the phrase, 'too many to fail.' Financial panics can be caused by herding and by contagion, as well as by big banks getting into trouble. In short, actively breaking up the largest banks would be a very complex task, with uncertain payoff." *NBER Reporter*, 2014, vol. 3. At http://www.nber.org/reporter/2014number3/2014number3.pdf.

Peter W. Culp, Robert Glennon, and Gary Libecap write about "Shopping for Water: How the Market Can Mitigate Water Shortages in the American West." "The American West has a long tradition of conflict over water. But after fifteen years of drought across the region, it is no longer simply conflict: it is crisis. . . . Many aspects of Western water law impose significant obstacles to water transactions that, given the substantial and diverse interests at stake, will take many years to reform. However, Western states can take an immediate step to enable more-flexible use of water resources by allowing simple, short-term water transactions. First, sensible water policy should allow someone who needs water to pay someone else to forgo her use of water or to invest in water conservation and, in return, to obtain access to the saved water. As a second step, state and local governments should facilitate these transactions by establishing essential market institutions, such as water banks, that can serve as brokers, clearinghouses, and facilitators of trade. Third, water managers should support and encourage the use of market-driven risk management strategies to address growing variability and uncertainty in water supplies. These strategies include the use of dry-year options to provide for water sharing in the face of shortages, and water trusts to protect environmental values. New reservoir management strategies that allow for sophisticated, market-driven use of storage could build additional resilience into water distribution. Fourth, states should better regulate the use of groundwater to ensure sustainability and to bring groundwater under the umbrella of water trading opportunities. Groundwater reserves are an important environmental resource and provide strategic reserves against drought, but proper management of groundwater is also critical to the development of markets. Markets cannot work effectively if users can delay facing the realities of local water scarcity through the unsustainable use of an open-access resource. Finally, strong federal leadership will be necessary to promote interstate and interagency cooperation in water management . . .” October 2014. At http://www.hamiltonproject.org/papers/shopping_for_water_how_the_market_can_mitigate_water_shortages_in_west/. Also see two other, related Hamilton Project Discussion Papers, also dated October 2014: Melissa S. Kearney, Benjamin H. Harris, Brad Hershbein, Elisa Jácome, and Gregory Nantz, “In Times of Drought: Nine Economic Facts about Water in the United States,” at http://www.hamiltonproject.org/papers/in_times_of_drought_nine_economic_facts_about_water_in_the_us/, and Newsha K. Ajami, Barton H. Thompson Jr., David G. Victor, “The Path to Water Innovation,” at http://www.hamiltonproject.org/papers/the_path_to_water_innovation/.

Stephen J. Ceci, Donna K. Ginther, Shulamit Kahn, and Wendy M. Williams address “Women in Academic Science: A Changing Landscape.” “We conclude by suggesting that although in the past, gender discrimination was an important cause of women’s underrepresentation in scientific academic careers, this claim has continued to be invoked after it has ceased being a valid cause of women’s underrepresentation in math-intensive fields. Consequently, current barriers to women’s full participation in mathematically intensive academic science fields are rooted in pre-college factors and the subsequent likelihood of majoring in these fields, and future research should focus on these barriers rather than misdirecting attention toward historical barriers that no longer account for women’s underrepresentation in academic science.” *Psychological Science in the Public Interest*, December 2014. At http://psi.sagepub.com/content/15/3/75.full.pdf+html.

E-books

Dirk Schoenmaker has edited *Macroprudentialism*, which includes 15 short and readable chapters from various perspectives. Here is a comment from Paul Tucker: “Legislators have typically favoured rules-based regulation. That is for good reason: it helps to guard against the exercise of arbitrary power by unelected officials. But a static rulebook is the meat and drink of regulatory arbitrage, which is endemic in finance. Finance is a 'shape-shifter'. That makes it hard to frame a regime that keeps risk-taking in the system as a whole within tolerable bounds. Instead, excessive risk-taking is likely to migrate to less regulated or unregulated parts of the system. . . . A number of implications for the design of macroprudential regimes flow from these features of the financial world. First, it will not be sufficient for bank regulation to be dynamically adjusted. It will also be necessary, for example, to vary minimum collateral (margin, haircut) requirements in derivatives and money markets when a cyclical upswing is morphing into exuberance; to tighten the regime applying to a corner of finance that is shifting from systemic irrelevance to a systemic threat; and to tighten the substantive standards, not only the disclosure standards, applying to the issuance of securities when the pattern of aggregate issuance is driving or facilitating excessive borrowing by firms or households. That means, second, that if finance remains free to innovate, adapt and reshape itself, every kind of financial regulator must be in the business of preserving stability. That needs to be incorporated into their statutory mandates and, more generally, into the design of regulatory agencies." A VoxEU.org eBook from the Duisenberg School of Finance and the Center for Economic Policy Press, 2014. http://www.voxeu.org/content/macroprudentialism.

Martin Neil Baily and John B. Taylor have edited *Across the Great Divide: New Perspectives on the Financial Crisis*, which includes 16 chapters and additional commentary from various authors. From their introduction: "The title is symbolic, first of all, of the range of different groups and opinions brought together, including, for example, those who have been harshly critical of the Federal Reserve Board and those who give high marks to the Fed's rescue efforts and unusual policy measures. In addition, while both Brookings and Hoover are proud of the range of scholars within each institution who embrace different politics and economic philosophies, Brookings is often seen as center left while Hoover is center right. So it was an important step to undertake this joint conference as a way of expanding the dialogue around monetary and regulatory policy. . . . This volume focuses on the 2008 financial crisis, the US response, and the lessons learned for future regulatory policy. . . . Part I of the book explains the causes and effects of the financial crisis. Part II focuses on the role played by the Federal Reserve before, during, and after the 2008 panic. Part III addresses the concept of 'too big to fail' (TBTF), and Part IV considers bankruptcy, bailout, and resolution." Hoover Institution Press, 2014. At http://www.hoover.org/research/across-great-divide-new-perspectives-financial-crisis-0.

About Economists

Jean Tirole has been awarded the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel for 2014 "for his analysis of market power and regulation." From the Nobel committee's "Information for the Public" essay: "In the

1980s, before Tirole published his first work, research into regulation was relatively sparse, mostly dealing with how the government can intervene and control pricing in the two extremes of monopoly and perfect competition. At this time, researchers and decision-makers were still looking for general principles that would apply to every industry. They advocated simple rules for regulatory policies, such as capping prices for monopolists and prohibiting cooperation between competitors in the same market, while permitting cooperation between firms at different positions in the value chain. Tirole's research would come to show that such rules work well in some conditions, but that they do more harm than good in others. Price caps can provide dominant firms with strong motives to reduce costs—a good thing for society—but may also permit excessive profits—a bad thing for society. Cooperation on price setting within a market is usually harmful, but cooperation regarding patent pools can benefit everyone involved. The merger of a firm and its supplier may lead to more rapid innovation, but it may also distort competition. To arrive at these results, a new theory was needed for oligopoly markets . . . Tirole's research would come to build upon new scientific methods, particularly in game theory and contract theory. . . . Game theory would aid the systematic study of how firms react to different conditions and to each other's behavior. The next step would be to propose appropriate regulation based on the new theory of incentive contracts between parties with different information. However, even though many people could see the research questions, they were difficult to solve." October 13, 2014. At http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2014/popular-economicsciences2014.pdf. The "Scientific Background" essay provides a fuller and more formal discussion of Tirole's work at http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2014/advanced-economicsciences2014.pdf.

Douglas Clement has interviewed Raj Chetty. Here's one interesting answer among many: "How has intergenerational mobility changed over time in America, and how does it vary across places within the U.S.? There's a popular conception that the U.S. once was a great land of opportunity and that that's no longer true today. Unfortunately, we've had relatively little data to actually be able to study the degree of social mobility systematically in the United States, so it is has been hard to know whether this conception is accurate or not. When we actually looked at the data over the past 30 to 40 years or so—a period for which we have good information from de-identified tax returns on children's parents' income as well as their own income—we find that, much to our surprise, there isn't that much of a difference in social mobility in the United States today relative to kids who were entering the labor force in, say, the 1970s or 1980s. That is, children's odds of moving up or down in the income distribution relative to their parents have not changed a whole lot in the past few decades. We find that where there is much more variation is across space rather than over time. . . . For example, for children growing up in places like Salt Lake City, Utah, or San Jose, California, the odds of moving from the bottom fifth of the national income distribution to the top fifth are more than 12 percent or even 14 percent in some cases, more than virtually any other developed country for which we have data. In contrast, in cities like Charlotte, North Carolina, Atlanta, Georgia, or Indianapolis, Indiana, a child's odds of moving from the bottom fifth to the top fifth are less than 5 percent—less than any developed country for which we currently have data." *The Region: Federal Reserve Bank of Minneapolis*, December 2014, pp. 10–24. At https://www.minneapolisfed.org/publications/the-region/interview-with-raj-chetty. Martin Feldstein contributed an overview of Chetty's work with "Raj Chetty: 2013 Clark Medal Recipient" in the Spring 2014 issue of this journal.

Renee Haltom has interviewed Nicholas Bloom. Here are some comments from Bloom on management quality and economic outcomes. "Economists have, in fact, long argued that management matters. Francis Walker, a founder and the first president of the American Economic Association, ran the 1870 U.S. census and then wrote an article in the first year of the *Quarterly Journal of Economics*, 'The Source of Business Profits.' He argued that management was the biggest driver of the huge differences in business performance that he observed across literally thousands of firms. Almost 150 years later, work looking at manufacturing plants shows a massive variation in business performance; the 90th percentile plant now has twice the total factor productivity of the 10th percentile plant. Similarly, there are massive spreads across countries—for example, U.S. productivity is about five times that of India. Despite the early attention on management by Francis Walker, the topic dropped down a bit in economics, I think because 'management' became a bad word in the field. Early on I used to joke that when I turned up at seminars people would see the 'M-word' in the seminar title and their view of my IQ was instantly minus 20. Then they'd hear the British accent, and I'd get 15 back." "I think the key driver of America's management leadership has been its big, open, and competitive markets. If Sam Walton had been based in Italy or in India, he would have five stores by now, probably called 'Sam Walton's Family Market.' Each one would have been managed by one of his sons or sons-in-law." "There's an old saying: What gets measured gets managed. I think in economics it's what gets measured gets researched. . . . Likewise with management—we hope if we can build a new multifirm and multicountry database, we can spur the development of the field." *Econ Focus: Federal Reserve Bank of Richmond*, Second Quarter, 2014, pp. 22–26. In this journal, Bloom wrote with John Van Reenen in the Winter 2010 about their research on "Why Do Management Practices Differ across Firms and Countries?" In the Spring 2014 issue of this journal, Bloom contributed a paper about "Fluctuations in Uncertainty," another subject discussed in the interview.

Complements of JEP

John Mueller and Mark G. Stewart discuss "Responsible Counterterrorism Policy." "[T]he United States spends about $100 billion per year seeking to deter, disrupt, or protect against domestic terrorism. If each saved life is valued at $14 million, it would be necessary for the counterterrorism measures to prevent or protect against between 6,000 and 7,000 terrorism deaths in the country each year, or twice that if the lower figure of $7 million for a saved life is applied. Those figures seem to be very high. The total number of people killed by terrorists within the United States is very small, and the number killed by Islamist extremist terrorists since 9/11 is 19, or fewer than 2 per year. . . . A defender of the spending might argue that the number is that low primarily because of the counterterrorism efforts. Others might find that to be a very considerable stretch. An instructive comparison might be made with the Los Angeles Police Department, which operates with a yearly budget of $1.3 billion. Considering only lives saved following the discussion above, that expenditure would be justified if the police saved some 185 lives every year when each saved life is valued at $7 million. (It makes sense to use the lower figure for the value of a saved life here, because police work is likely to have few indirect and ancillary costs: for example, a fatal car crash does not cause others to avoid driving.) At present, some 300 homicides occur each year in the city and about the same number of deaths from automobile accidents. It is certainly plausible to suggest that both of those numbers would be substantially higher without police efforts, and accordingly that local taxpayers are getting pretty good value for their money. Moreover, the police provide a great many other services (or 'cobenefits') to the community for the same expenditure, from directing traffic to arresting burglars and shoplifters." Cato Institute Policy Analysis, Number 755, September 10, 2014. At http://object.cato.org/sites/cato.org/files/pubs/pdf/pa755.pdf. This article complements the article by the same authors on "Evaluating Counterterrorism Spending" in the Summer 2014 issue of this journal.

Tomasz Koźluk and Vera Zipperer have published "Environmental Policies and Productivity Growth: A Critical Review of Empirical Findings." "The traditional approach sees environmental policies as a burden on economic activity, at least in the short to medium term, as they raise costs without increasing output and restrict the set of production technologies and outputs. At the same time, the Porter Hypothesis claims that well-designed environmental policies can provide a 'free lunch'—encouraging innovation, bringing about gains in profitability and productivity that can outweigh the costs of the policy. This paper reviews the empirical evidence on the link between environmental policy stringency and productivity growth. . . . [M]any of the studies are fragile and context-specific . . . Practical problems . . . include: improving the measurement of environmental policy stringency; investigating effects of different types of instruments and details of instrument design; exploiting cross-country variation; and the complementary use of different levels of aggregation." *OECD Journal: Economic Studies*, 2014, vol. 1, pp. 1–32. Can be read at http://www.oecd-ilibrary.org/economics/environmental-policies-and-productivity-growth_eco_studies-2014-5jz2drqml75j. For an early statement of the Porter hypothesis and a counterpoint, in the Fall 1995 issue of this journal Michael E. Porter and Claas van der Linde wrote "Toward a New Conception of the Environment-Competitiveness Relationship," and Karen Palmer, Wallace E. Oates, and Paul R. Portney respond with "Tightening Environmental Standards: The Benefit-Cost or the No-Cost Paradigm?"

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.

Fair Trade Coffee

Fair trade coffee is a cup half full, according to Raluca Dragusanu, Daniele Giovannucci, and Nathan Nunn in "The Economics of Fair Trade" (Summer 2014, vol. 28, no. 3, pp. 217–36). We are not persuaded.

The authors barely mention the fees imposed upon current and prospective fair trade coffee growers by FLO-CERT, the organization that verifies and certifies fair trade products. By not spelling out the fees, the authors may leave readers with a mistaken impression that the fees are trifling. Elliott (2012) summarizes nicely the latest fee structure. For cooperatives of poor producers, the initial application fee is €525, and fees for the first inspection vary from €1430 to €3470 depending upon a cooperative's size. While certifications are good for three years, annual fees range from €1170 to €2770 and include interim surveillance of growers' practices. In short, Fairtrade International requires farmers in low-income countries to pay thousands of dollars in order to participate in a network presumably intended to offer poverty relief to its producer organizations as well as protection from allegedly ruthless local monopsonist coffee buyers, called *coyotes*. The existence of these large and explicit costs to growers casts some doubt on the relatively optimistic conclusions of this paper.

As the authors acknowledge, only a small fraction of coffee grown by fair trade producers is able to be sold as fair trade coffee, but readers should also be clear that applying to join the fair trade network does not guarantee a willing buyer on the other side of the market. As Fridell (2007) notes, newcomers to fair trade production are the least likely to benefit because they cannot compete on an equal footing with established cooperatives in an already saturated market. Fridell cites Martinez (2002), who in turn describes the plight of a certified producer organization that searched for eight years to locate a willing buyer.

Fair trade also appears to exacerbate inequality in certain ways. Since 2007, the minimum price guaranteed by Fairtrade International has rarely been binding: coffee prices have been high, so all coffee growers—fair trade or not—have been receiving the market price for their crops. In this case the primary benefit to fair trade growers is the social premium—currently 20 cents per pound. Valkila (2014) observes that should prices plummet as they did during the 1990s, the fair trade growers who would benefit most from the minimum-price guarantee are those already poised to supply significant quantities of high-quality coffee to the network. According to Valkila, these growers are likely male owners of large tracts of land and most likely grow coffee in locations where the standard of living is already relatively high. For example, according to Fairtrade Foundation (2012) data, in 2009–2010, Peru supplied 25 percent of all fair trade coffee, while Tanzania was the tenth-largest fair trade coffee supplier, supplying merely 4 percent of the total. Coffee producers in Peru are clearly better-positioned to benefit from fair trade than those in Tanzania. But in 2009–2010, Peru had a per capita GDP of roughly $4,500 (in current dollars, World Bank data), while Tanzania had a per capita GDP of just over $500.

Fair trade also has some other little-known distributional consequences. For example, a growing literature suggests that the benefits of fair trade coffee accrue mainly to those in the supply chain who are already well-off by global standards.

Valkila, Haaparanta, and Niemi (2010) assess the distribution of benefits from fair trade between the producing and consuming nations. Tracing coffee grown in Nicaragua and consumed in Finland, Valkila et al. discover that a larger fraction of the retail price of fair-trade-labeled coffee remains in the consuming country relative to the case of conventionally marketed coffee. One can debate the extent to which coffee growers in low-income countries are enriched via fair trade, but there is little doubt that traders and roasters benefit immensely. Sylla (2014), a former fair trade insider, provides a book-length treatment of such distributional consequences.

A number of other points could be made. For example, the authors did not mention the research which assesses fair trade as price discrimination between more-caring and less-caring consumers—in which those who care more end up paying more. We were intrigued that the authors cite Beuchelt and Zeller (2011)—popularly known as "the Hohenheim study"—while omitting its primary finding: after ten years of participation in fair trade networks, Nicaraguan fair trade coffee growers grew poorer relative to their conventional counterparts.

In the close of the paper, the authors claim that even if fair trade is not very effective, it nevertheless constitutes an improvement over direct aid efforts because direct aid efforts are subject to misuse by corrupt dictators and bureaucrats. We agree wholeheartedly that *governmental* aid efforts with little accountability breed corruption; Leeson and Sobel (2008) demonstrate that this truth holds even when aid is transferred between governments located within the national boundaries of a developed country. Yet this conclusion—that fair trade trumps direct aid because of the likelihood of corruption with direct aid—discounts an obvious third way to improve the lives of the poor, and ignores an abiding truth of economic growth and development. When poor people everywhere raise their incomes in a lasting way, it typically happens because these people are able to accumulate superior physical and human capital. A wide range of low-cost and simple human capital investments in nutrition and in basic language and math skills can alter the plight of those who remain in extreme global poverty. We propose that caring coffee consumers should not waste their money purchasing overpriced coffee that enriches mainly importers, roasters, retailers, and fair trade bureaucrats—in short, everyone but the poor it claims to help. We suggest that the global poor achieve greater, more enduring gains when we purchase the coffee we like most and that fits our budgets best, and entrust our charitable giving not with the fair trade network, but instead with the nongovernment organizations best-positioned to knowledgably and effectively invest directly in the human beings we want most to help.

Victor V. Claar
Henderson State University
Arkadelphia, Arkansas

Colleen E. Haight
San Jose State University
San Jose, California

References

Beuchelt, Tina D., and Manfred Zeller. 2011. "Profits and Poverty: Certification's Troubled Link for Nicaragua's Organic and Fairtrade Coffee Producers." *Ecological Economics* 70(7): 1316–24.

Elliott, Kimberly. 2012. "Is My Fair Trade Coffee Really Fair? Trends and Challenges in Fair Trade Certification." Center for Global Development, CGD Policy Paper 017, December.

Fairtrade Foundation. 2012. "Fairtrade and Coffee." Commodity Briefing, May. http://www.fairtrade.net/fileadmin/user_upload/content/2009/resources/2012_Fairtrade_and_coffee_Briefing.pdf.

Fridell, Gavin. 2007. *Fair Trade Coffee: The Prospects and Pitfalls of Market-Driven Social Justice.* University of Toronto Press.

Leeson, Peter T., and Russell S. Sobel. 2008. "Weathering Corruption." *Journal of Law and Economics* 51(4): 667–81.

Martinez, Maria Elena. 2002. "Poverty Alleviation through Participation in Fair Trade Coffee Networks: The Case of the Tzotzilotic Tzobolotic Coffee Coop, Chiapas, Mexico." September. http://welcome2.libarts.colostate.edu/centers/cfat/wp-content/uploads/2009/09/Case-Study-Tzotzilotic-Tzobolotic-Coffee-Coop.pdf.

Sylla, Ndongo S. 2014. *The Fair Trade Scandal: Marketing Poverty to Benefit the Rich.* Athens, OH: Ohio University Press.

Valkila, Joni. 2014. "Do Fair Trade Pricing Policies Reduce Inequalities in Coffee Production and Trade?" *Development Policy Review* 32(4): 475–93.

Valkila, Joni, Pertti Haaparanta, and Niina Niemi. 2010. "Empowering Coffee Traders? The Coffee Value Chain from Nicaraguan Fair Trade Farmers to Finnish Consumers." *Journal of Business Ethics* 97(2): 257–70.

Journal of Economic Perspectives—Volume 29, Number 1—Winter 2015—Pages 217–220

Editorial Note
Correction to Richard S. Tol's "The Economic Effects of Climate Change"

In the Spring 2009 issue, this journal published "The Economic Effects of Climate Change" by Richard S. J. Tol (vol. 23, no. 2, pp. 29–51). The paper included a figure summarizing the results of a number of studies, showing their estimates of how the economic costs of climate change varied with the predicted change in global temperatures. In early 2014, the editors received a complaint pointing out errors in the paper: specifically, several estimates had not been accurately transferred from the original studies. In the Spring 2014 issue, we published a "Correction and Update: The Economic Effects of Climate Change" (vol. 28, no. 2, pp. 221–26) by Richard Tol. However, this version also contained errors that were soon pointed out by various researchers. The editors discussed the situation with Richard Tol and with outside reviewers at some length.

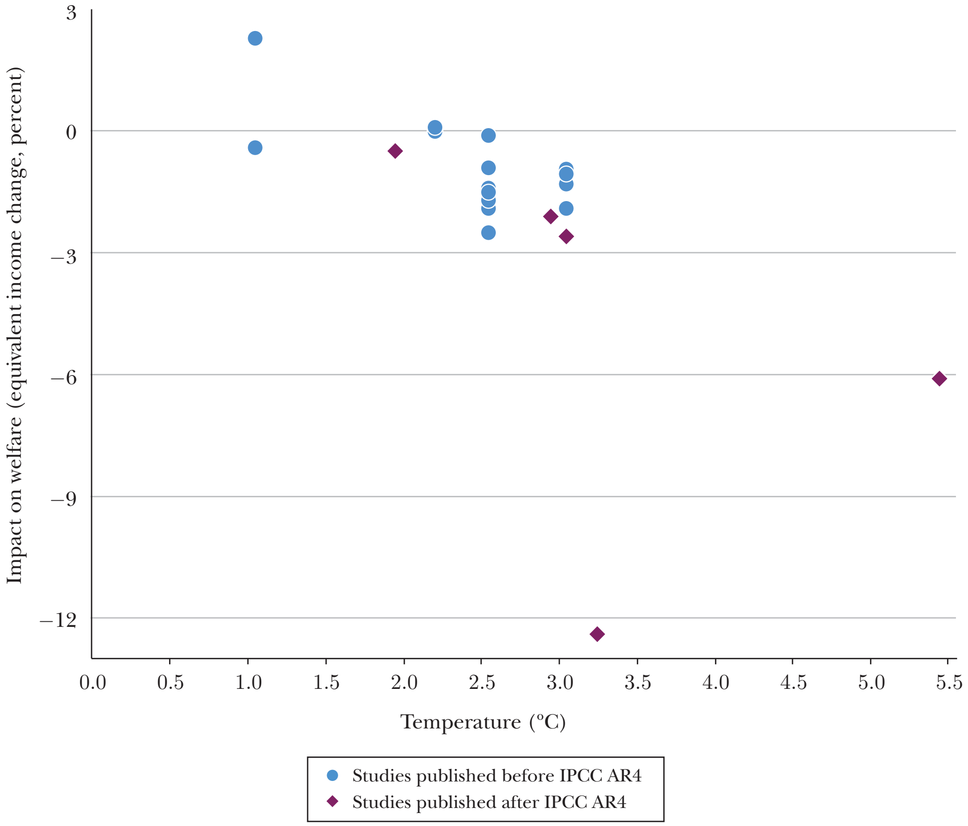
This correction offers a final revision and update to the figure in question. This figure is republished from the most recent report of the International Panel on Climate Change (IPCC), in Chapter 10 of the volume *Climate Change 2014: Impacts, Adaptation, and Vulnerability*. Richard Tol is one of the two "Coordinating Lead Authors" of this chapter, along with Douglas J. Arent, but the chapter also draws on efforts by a group of other lead authors, contributing authors, and review editors. Figure 1 reproduces Figure 10-1 from p. 690 of the IPCC report. The IPCC discussion of this figure offers some useful cautions about interpretation:

> Estimates agree on the size of the impact (small relative to economic growth), and 17 of the 20 impact estimates shown in Figure 10-1 are negative. Losses accelerate with greater warming, and estimates diverge. The new estimates have slightly widened the uncertainty about the economic impacts of climate. Welfare impacts have been estimated with different methods, ranging from expert elicitation to econometric studies and simulation models. Different studies include different aspects of the impacts of climate change, but no

Figure 1

The Economic Costs of Changes in Global Temperatures

Figure 10-1 | Estimates of the total impact of climate change plotted against the assumed climate change (proxied by the increase in the global mean surface air temperature); studies published since IPCC AR4 are highlighted as diamonds; see Table SM10-1.



Notes: IPCC AR4 refers to the Fourth Assessment report of the IPCC, which was published in 2007. This figure is from AR5, the Fifth Assessment report of the IPCC, published in 2014.

estimate is complete; most experts speculate that excluded impacts are on balance negative. Estimates across the studies reflect different assumptions about inter-sectoral, inter-regional, and inter-temporal interactions, about adaptation, and about the monetary values of impacts. Aggregate estimates of costs mask significant differences in impacts across sectors, regions, countries, and populations. Relative to their income, economic impacts are higher for poorer people.

The original figure in the 2009 JEP article estimated a best-fit line passing through the points on this kind of figure, along with confidence intervals for that estimate. Given the differences across the studies as mentioned in the IPCC report, several outside reviewers involved in our editorial process expressed a concern that such estimates were not meaningful. As shown, the figure in the IPCC report does not seek to estimate a best-fit line or confidence intervals, but only offers a summary of the results from existing studies. Tol offers further discussion of the curve-fitting issues with this kind of data in "Bootstraps for Meta-Analysis with an Application to the Impact of Climate Change," forthcoming in *Computational Economics* (doi: 10.1007/s10614-014-9448-5).

For a list of studies that were included, what methods were used in these studies, what economic sectors were covered, and the like, we would point interested readers to the "Supplementary Material" table for Chapter 10. The full report, Chapter 10, and the Supplementary Material are all available at http://www.ipcc.ch/report/ar5/wg2/. Controversy over these estimates seems likely to continue. We recommend that readers interested in these questions use the figure and data from the IPCC report as their starting point.

Call for Sessions and Papers for the January 2016 American Economic Association Annual Meeting

Members wishing to give papers or organize complete sessions for the program for the meetings in Boston are invited to submit proposals electronically to Professor Robert Shiller via the American Economic Association website ***starting on March 1***. While papers covering a wide array of topics in economics will be included on the 2016 program, Professor Shiller especially encourages proposals that cross the boundaries of conventionally-defined disciplines.

To be considered, individual paper proposals (with abstracts) and up to *two Journal of Economic Literature bibliographic codes in rank order should be submitted* **by April 1, 2015**. The deadline for complete session proposals is **April 15, 2015**. *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 (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). Papers on econometric or mathematical methods 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.

AMERICAN ECONOMIC ASSOCIATION

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.

MIX
Paper from responsible sources
FSC www.fsc.org
FSC™ C101537

The Journal of

Economic Perspectives

Winter 2015, Volume 29, Number 1

AMERICAN ECONOMIC ASSOCIATION