The Journal of Economic Perspectives

A journal of the American Economic Association

Summer 2017

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 LSC Communications, Owensville, Missouri, 65066, USA.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Summer 2017, Vol. 31, No. 3. The *JEP* is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $20.00, $30.00, or $40.00 depending on income; for an additional $15.00, you can receive this journal in print. E-reader versions are free. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 31 • Number 3 • Summer 2017

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 31, Number 3—Summer 2017—Pages 3–28

International Monetary Relations: Taking Finance Seriously

Maurice Obstfeld and Alan M. Taylor

The architecture of the international monetary and financial system is a major determinant of how close the world economy can come to realizing its potential, and how serious are the risks of crisis and disruption. In this essay, we particularly want to highlight the interactions of the international monetary system with financial conditions, and not just with the output, inflation, and balance of payments goals that have been central to most accounts.

A basic constraint on the design of all international monetary systems is the *monetary policy trilemma*: a country can enjoy two of the following three features simultaneously, but not all three: exchange-rate stability, freedom of cross-border payments, and a primary orientation of monetary policy toward domestic goals (for example, Keynes 1930, chap. 36; Padoa-Schioppa 1988; Obstfeld and Taylor 1998, 2004; for a brief intellectual history, see Irwin 2011). For more than a century, efforts to cope with the monetary trilemma have varied across time and space, with mixed success. For example, the gold standard of the late 19th and early 20th centuries implied fixed exchange rates because all gold-standard central banks fixed their currencies' values in terms of gold. Coupled with international capital mobility, however, the gold standard meant that autonomous monetary policy was infeasible. Conversely, the Bretton Woods system that operated from the end of World War II

■ *Maurice Obstfeld is Economic Counsellor and Director of Research at the International Monetary Fund, Washington, DC. He is on leave as Class of 1958 Professor of Economics, University of California Berkeley, Berkeley, California. Alan M. Taylor is Professor of Economics and Finance, University of California Davis, Davis, California.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.3 doi=10.1257/jep.31.3.3

into the early 1970s mandated fixed exchange rates but, for as long as international capital mobility was blocked, countries could, to some degree, use monetary policy for domestic goals. In recent decades, many advanced economies have moved to a system of floating exchange rates: in the context of the monetary trilemma, their tradeoff was to sacrifice fixed exchange rates in order to allow both international capital mobility and a monetary policy geared toward domestic objectives.

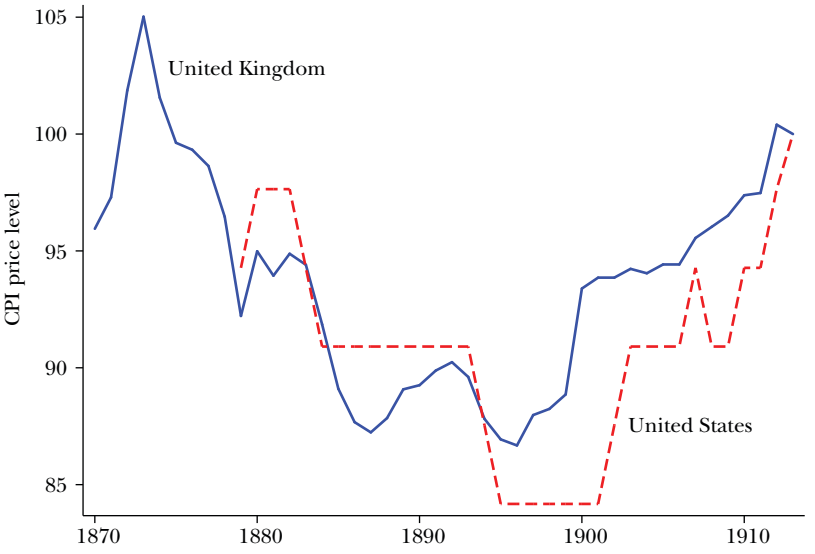
While the monetary trilemma is a useful organizing principle for categorizing different choices about international monetary systems, we need to be clear that it does not imply that one choice is the best, much less that any choice can solve all economic problems or insulate an economy fully from foreign financial disturbances. In this essay, we review how financial conditions and outright financial crises have posed difficulties for each of the main international monetary systems in the last 150 years or so: the gold standard, the interwar period, the Bretton Woods system, and the current system of floating exchange rates. We will argue that the Bretton Woods agreement of 1944 addressed only a limited set of issues, those most relevant after the traumatic transformations of the Great Depression and World War II, which included a marked retrenchment in national and international financial market activities. However, a broader set of financial stability challenges was not addressed at Bretton Woods. Always latent, these dangers had periodically exploded into central importance in the world economy from the 1870s to the 1930s—and, despite a long period in abeyance after World War II, they would gradually take on increasing importance as the postwar decades passed. Indeed, considering the distinct policy challenges in this dimension, a *financial trilemma* has been proposed to complement the better-known monetary trilemma: specifically, countries must choose among national sovereignty over financial stability policy, integration into global financial markets, or financial stability—but they cannot have all three (Schoenmaker 2013).

Our essay will rest on the argument that—even as the world economy has evolved and sentiments have shifted among widely different policy regimes—three fundamental challenges for any international monetary and financial system have remained. How should exchange rates between national currencies be determined? How can countries with balance of payments deficits reduce these without sharply contracting their economies and with minimal risk of possible negative spillovers abroad? How can the international system ensure that countries have access to an adequate supply of international liquidity—financial resources generally acceptable to foreigners in all circumstances? In concluding, we evaluate how the current international monetary system answers these questions.

The Bretton Woods Regime and Its Contradictions

The slide into World War II led to effective financial autarky for many countries. The immediate postwar years then saw widespread tightening of government's grips over banks and financial markets (Cassis 2011, pp. 108–9). More generally, *laissez faire* ideology was in retreat (Polanyi 1944), and unregulated financial markets drew special

Figure 1
Price Levels under the Gold Standard, United Kingdom 1870–1913 and United States 1870–1913



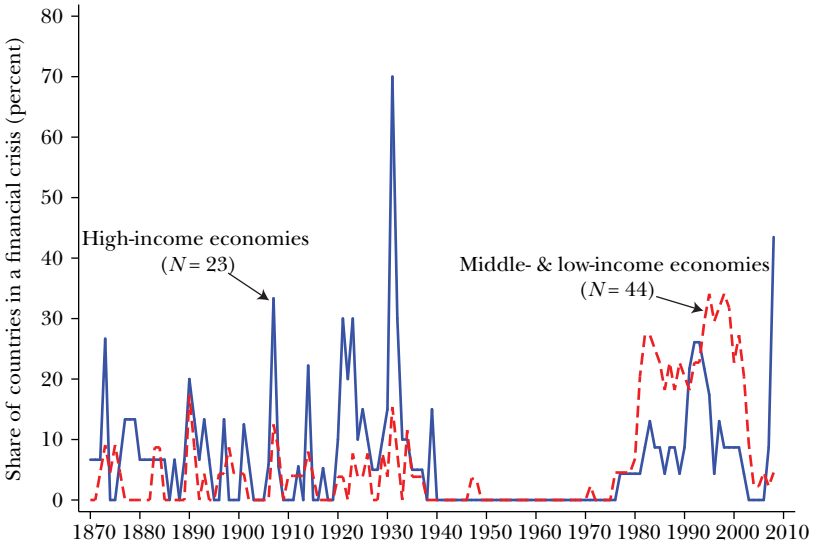
Source: Data from Jordà, Schularick, and Taylor (forthcoming) Macrofinancial Database.

opprobrium, as they were widely perceived to have failed. The underlying premise of the 1944 Bretton Woods conference was that neither the classical gold standard nor the successor arrangements during the interwar period had worked well.

Historical Context: The Gold Standard

Under the pre-1914 gold standard, the monetary trilemma was resolved in favor of exchange stability and freedom of foreign transactions. While these features did tend to promote an expansion of trade and international lending, the system severely limited the role monetary policy could potentially play in macroeconomic stabilization. Short-term interest rates in different countries tracked each other relatively closely (Obstfeld, Shambaugh, and Taylor 2005). At the same time, longer-term inflation trends were shared across countries and tied to supply and demand forces in the global gold market. Thus, price levels under the gold standard sometimes underwent long periods of decline or increase as shown in Figure 1, generally falling from about 1880 to 1895 in the face of limited gold supplies, then rising through 1914 in response to gold discoveries in the Yukon and South Africa. These long swings in prices could cause tensions, both economic and political, and countries had to cope with unanticipated redistributions between paper debtors and creditors. Notably, the stability of banks and the financial system was not assured by gold convertibility of currency, as evidenced by the 19th century history of banking

Figure 2
Financial Crises, 1870–Present



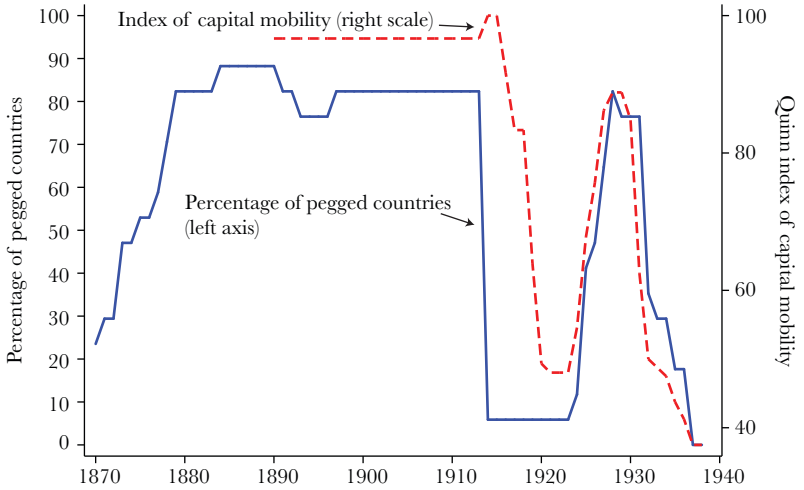
Source: Data from Qian, Reinhart, and Rogoff (2011).

crises both in the United States (Jalil 2015) and elsewhere.[1] Figure 2 shows the pattern of financial crises affecting advanced economies since 1870.

Around the same time as the Panic of 1873, to focus on one prominent episode of financial crisis, Bagehot's (1873) *Lombard Street* famously laid out the Bank of England's role as the financial markets' *lender of last resort* (although this role had been described earlier by Thornton in his 1802 masterpiece, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*). Bagehot's advice was that a central bank during a financial panic should lend freely against good collateral. But how could the central bank increase the money supply in this way while simultaneously maintaining its currency's parity with gold? When confronted with both a banking and a currency crisis, Bagehot (1873) viewed maintaining the gold standard as the priority: "We must look first to the foreign drain, and raise the rate of interest as high as may be necessary. Unless you stop the foreign export you cannot allay the domestic alarm" (pp. 27–28). Bagehot's argument amounted to the assertion that monetary policy could be deployed to stem a banking panic independent of the exchange-rate constraint, which might be true in certain special circumstances, but more broadly serves to illustrate how some resolutions of the monetary trilemma could simultaneously exacerbate financial instability.[2] In another episode, in 1907

[1] For overviews of the macroeconomics of the pre–World War I gold standard, useful starting points are Cooper (1982), Bordo and Schwartz (1984), and Eichengreen (2008, chap. 2).
[2] See Laidler (2003) on the contrasting views of Bagehot and Thornton regarding the relative importance of internal versus external stability.

Figure 3
Pegging to Gold and Capital Mobility, 1870–1938



Source: Data from Jordà, Schularick, and Taylor (forthcoming) Macrofinancial Database; Quinn, Schindler, and Toyoda (2011).

the Bank of England, alarmed by gold outflows that financed overheating financial markets in the United States, abruptly hiked its target interest rate, helping to set off the devastating panic of 1907.

Though the 1873 and 1907 episodes are among the better-known ones, they are merely two of the many severe systemic banking crises and accompanying severe recessions, sometimes occurring at once in several countries, that punctuated the gold standard era. Indeed, the panic of 1873, which afflicted Europe as well as North America, helped inspire the founding of the German Reichsbank in 1876, while the US panic of 1907, against the backdrop of periodic liquidity tensions in the US banking system, led to the founding of the US Federal Reserve.

Historical Context: The Interwar Period

World War I surpassed previous wars not only in its scope and destructiveness, but also in the extent to which economic relationships between nations broke down. That breakdown was in part a result of direct government actions, including widespread suspension of the gold standard and, significantly, pervasive official control over external payments, a huge contrast to the previous era's *laissez faire*. Looking back, Keynes, who had served in the UK Treasury during the war, said, "Complete control was so much against the spirit of the age, that I doubt it ever occurred to any of us that it was possible" (as cited in Obstfeld and Taylor 2004, p. 146). Governments had opened Pandora's box.

Figure 3 illustrates the pattern that followed. The 1920s saw various attempts by governments to remove exchange control and return to gold: only about 10 percent

of currencies were still pegged to gold in the early 1920s, but by the end of the 1920s, 80 percent were again pegged to gold, and capital mobility was once more widespread. In the subsequent Great Depression, most countries abandoned the gold standard and imposed harsh capital controls (Obstfeld and Taylor 2004, pp. 136–40).

The story of the Great Depression from a US perspective is well known. The US economy succumbed to macroeconomic and financial shocks as US government policy failed to react effectively. Waves of banking crises followed, a pattern seen in many countries around the world in the 1930s, as reflected in Figure 2 presented earlier. The Federal Reserve failed to do much as a lender of last resort, despite having been founded to fill that role (Hetzel and Richardson 2016). Thus, various historians have attributed the depth of the Great Depression in the US to Federal Reserve incompetence (Friedman and Schwartz 1963; Hsieh and Romer 2006), or a collapse of credit (Mishkin 1978; Bernanke 1983; Bernanke and James 1991), or both. Taken together, these arguments indicate the importance of both traditional macroeconomic and financial factors.

However, the economic and financial crisis of the Great Depression also occurred within an international context, driven in part by problems that arose from the global attempt to return to the gold standard (Temin 1989; Eichengreen 1992). The United Kingdom returned to gold in 1925, but at the prewar sterling–gold parity, despite a significantly higher postwar price level compared with 1913. France returned in 1926, but could tolerate doing so only at a much-depreciated exchange rate between the franc and gold. These fateful decisions ensured that for many years the deflated British economy would struggle with a strong currency, high unemployment, and gold losses (Keynes 1925); in contrast, reflated France enjoyed a weak currency and a gold surplus (Hamilton 1988; Irwin 2012–2013). Contradicting textbook stories about price-specie-flow adjustment, these outcomes highlighted the real-world asymmetry between deficit countries, who were pressured by balance of payments outflows, and surplus countries, who faced no corresponding pressure to reduce their external imbalances.

The United States, which had remained on gold throughout World War I and after, was by the late 1920s experiencing a massive stock market boom that attracted substantial gold inflows from abroad (Kindleberger 1973 [2013]). US credit tightening compounded the inflow, and countries throughout the world raised interest rates as they competed to retain gold. This purposeful competition for gold ultimately proved deflationary, and escape came slowly. Britain abandoned the gold standard in 1931. Other countries followed. Instability in global banking played an important role in driving speculative capital flows (Borio, James, and Shin 2014). In the US economy, the first signs of economic stabilization occurred only in spring 1933, when President Roosevelt also suspended the US dollar's gold link. The end of tight money stopped the collapse of price levels and nurtured hesitant recoveries in countries that depreciated (Eichengreen and Sachs 1985; Campa 1990; Bernanke and Carey 1996; Obstfeld and Taylor 1998).

From the perspective of this essay, two lasting legacies of this period are worth emphasizing. One was a fear of "beggar-thy-neighbor" policies, a phrase originally due to Adam Smith,[3] but now widely linked with the Depression era. Countries tried in several ways to bottle in domestic demand at the expense of their trading partners, including high tariff walls and strict exchange controls. Competitive currency depreciation was also often held up as a poster child in this policy class, its typical goal being to switch demand between countries (for example, League of Nations 1944). As Eichengreen and Sachs (1985) pointed out, however, simultaneous competitive monetary expansion in a group of countries where each one is trying to depreciate, even if it leaves their currencies' mutual exchange rates unchanged, could be a better equilibrium if all are battling deflation and unemployment.

The other major legacy was that the financial instability of the interwar period left governments much less willing to tolerate free-wheeling financial markets. In the United States in 1933, for example, this new mindset begat the Glass–Steagall act, which prohibited commercial banks from engaging in investment-banking activities; the creation of the Federal Deposit Insurance Corporation, to oversee a new system of deposit insurance; new and broad regulatory powers for the Federal Reserve; and Regulation Q, which imposed interest-rate ceilings to discourage banks from competing for deposits. The Securities Exchange Act of 1934 and the Banking Act of 1935 soon followed.

On economic policy, doctrinal change was swift and dramatic. Macroeconomic policy was seen to have been badly wrong. By the 1940s, new thinking, as represented by Keynes and his followers, was the order of the day. There would be no rush to restore either a gold standard or unregulated financial markets, as there had been after World War I. As Cassis (2011) describes it, the turbulent years from 1914 to 1945 "led to an ideological shift which, combined with a generational change, favored state intervention and a more organized form of capitalism" (p. 109). This was the ascendant worldview as international negotiators gathered at Bretton Woods, New Hampshire, in July 1944 to design the postwar international monetary and financial order.

The Bretton Woods Approach and the Creation of the IMF

Post–World War II reconstruction offered an opportunity to construct a new international monetary system. Ruggie (1982) painted the contrast between earlier attitudes and the new postwar vision of this system: "[U]nlike the economic nationalism of the thirties, it would be multilateral in character; unlike the liberalism of

[3]Smith (1776) wrote in *The Wealth of Nations* (Book IV, Chapter III): "[N]ations have been taught that their interest consisted in beggaring all their neighbours. Each nation has been made to look with an invidious eye upon the prosperity of all the nations with which it trades, and to consider their gain as its own loss. Commerce, which ought naturally to be, among nations, as among individuals, a bond of union and friendship, has become the most fertile source of discord and animosity." Any similarity with current political discourse is not in the least coincidental.

the gold standard and free trade, its multilateralism would be predicated upon domestic interventionism."

Under the system designed at Bretton Woods in 1944, exchange rates were fixed, with every country pegging to the US dollar (and thereby stabilizing the $N-1$ exchange rates among the N currencies), while the United States was supposed to peg the dollar price of gold (an arrangement that formally applied mainly to its transactions with official foreign dollar holders, and thus gave the US in practice an asymmetrically central position with disproportionate power over global monetary conditions). Unlike the euro-area monetary union of recent times, the Bretton Woods system mandated no external constraints on government budgets, allowing fiscal policy to be used more freely as a tool of macro stabilization.

With the recognition that countries with fixed exchange rates might run short of international reserves, the International Monetary Fund was created as an emergency lender. Countries also had the capacity, subject to IMF approval, to devalue or revalue their currencies in circumstances of "fundamental disequilibrium"—a term nowhere defined in the IMF's Articles of Agreement. The basic idea was that countries running *persistent* balance of payments deficits should not be forced to maintain what appeared to be an unsustainably strong exchange rate through employment-reducing monetary contraction, fiscal austerity, or both. Rather, as Keynes put it in defending the plan before the British Parliament, the value of the currency would adjust to the economy's needs, not the reverse.

Of course, in oxymoronic fashion, "fixed but adjustable" exchange parities do face the frequent drawback that markets can often see the changes coming—or imagine that they will come—and in those cases, speculative capital flows (self-fulfilling or anticipatory) can disrupt any pretense of deliberate and consultative exchange-rate adjustment. The problem was well understood from the interwar experience, but the risks were mitigated when the IMF opened its doors in 1946: pervasive capital and exchange controls remained and domestic financial systems were broadly constrained and repressed, greatly reducing crisis risk and limiting speculative responses to possible exchange parity changes.[4] Nor did the IMF's Articles have as a goal any process of capital-control liberalization. Indeed, Article VI, Section 1(a), discouraged members from using IMF resources to finance sustained capital flight, and also allowed the IMF to request a member to impose outflow controls in such cases. Article VI, Section 3, explicitly stated, "Members may exercise such controls as are necessary to regulate international capital movements," subject to some restrictions. Deviations from frictionless capital mobility, to greater of lesser extent, then gave national authorities scope to manage domestic interest rates notwithstanding fixed exchange rates.

In sum, by eliminating capital mobility, the Bretton Woods system set up a resolution of the monetary trilemma based on exchange-rate stability and a degree of autonomy of monetary policy. Long-run inflation trends would be determined *de facto*

[4]The exception that proved the rule was Britain in 1947, where a premature return to free sterling convertibility quickly ended in a balance of payments crisis and a return to capital controls.

by US monetary policy (mediated by the nature of the dollar's link to gold); but *in extremis*, countries could also adjust currency values. IMF funding was meant to ensure that such adjustments would occur only in response to highly persistent shocks.

The IMF's Articles did not explicitly address financial market stability. But in the absence of extensive private international capital flows, each country had a free hand to regulate its financial sector. With memories of the 1930s still fresh, an inclination for tight regulation, coupled with relatively uncomplicated financial systems, made the Bretton Woods period up to about 1970 almost crisis-free compared with the decades that preceded and followed it. Figure 2 illustrates how singular that interlude was.

These postwar choices mirrored deeper economic objectives. According to the first of its Articles of Agreement, two of the IMF's original main purposes are to "facilitate the expansion and balanced growth of international trade" and "to assist in the establishment of a multilateral system of payments in respect of current transactions between members and in the elimination of foreign exchange restrictions which hamper the growth of world trade." Through the start of the 1960s, these goals seemed to have been well realized—supported also by the Marshall Plan after World War II, five tariff reduction rounds under the GATT, falling international transport costs, as well as the European Payments Union and a range of other European integration initiatives. International trade did recover, but weaknesses in the Bretton Woods architecture were lurking.

Weaknesses Emerge in the Bretton Woods Architecture

First, stability of the Bretton Woods fixed exchange rates was predicated on continuing limited cross-border capital mobility. Policymakers struggled with financial plumbing, trying to open the pipes for payments on current transactions to support the rebirth of global trade but close the valves for speculative capital transactions that could destabilize the system. However, one result of the Bretton Woods system's successes was that the opportunity for capital flows inevitably grew and unwanted leakages increasingly seeped through. Fixed exchange rates therefore became harder to maintain. As early as 1961, the German and Dutch currencies were revalued in the face of large capital inflows. This episode was a harbinger of much bigger eruptions later in the 1960s, notably the devaluations by Britain and several others in November 1967.

Second, and parallel with the increased exchange-rate instability implied by greater capital mobility, was a phenomenon that remains central to financial-stability policy to this day: the migration of financial activity to less-regulated venues, both through the location of banking activity offshore and domestic financial innovation. In the 1960s, US banks were constrained by the Depression-era Regulation Q from competing for deposits onshore; but moving offshore—notably to London, where they could operate essentially free of regulation—allowed them to circumvent the rule. In addition, the 1963 Interest Equalization Tax, intended to strengthen the US balance of payments by taxing capital outflows, gave multinational firms an incentive to borrow dollars from foreign banks and issue dollar bonds abroad. Eurodollar and

eurodollar-bond markets arose in London, ultimately helping London to become the world's pre-eminent financial center. Because of the Regulation Q interest-rate ceilings, mounting US inflation also implied that the real interest rates banks could offer savers were becoming increasingly negative. As financial activity moved to commercial paper markets and new money-market mutual funds, pressures for bank deregulation grew in the United States as well as in other industrial countries.[5]

A third weakness in the Bretton Woods system centered on international liquidity. Governments around the world were accumulating US dollars to hold as international reserves, while the United States had promised to redeem foreign official dollars at a price of $35 per ounce. All would be well as long as the Americans held enough gold, but as Triffin (1960) pointed out, redemption would become increasingly problematic as global dollar reserves in foreign hands continued to grow. In 1960, foreign US dollar reserves overtook the value of US gold holdings and speculators began to push up gold's price in the London market, which raised the possibility that the United States (like the Bank of England in the 1930s), might have trouble meeting official demands to convert its currency into gold.

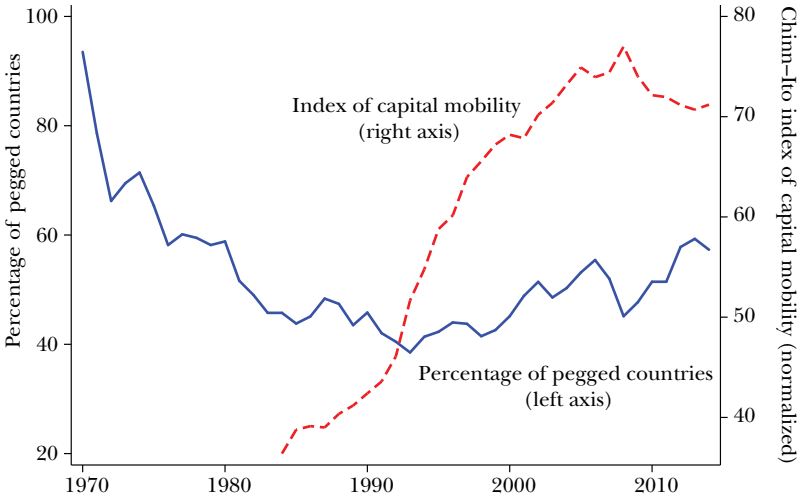
US inflation began to rise in the latter 1960s. The supposed link of the US dollar to gold had weakened significantly over time, causing a problem for countries pegged to the dollar as they faced pressure to import inflation from the United States. At the same time, analysts and markets began to believe that the US dollar was overvalued and in need of depreciation. The resulting capital flows into currencies like Germany's mark, Japan's yen, and Switzerland's franc exacerbated the inflationary pressures those countries faced, as their central banks had to buy dollars to keep their exchange rates pegged, in the process increasing their international reserve holdings and money supplies, as well as their exposure to any action by the United States to increase the dollar price of gold. Triffin's feared imbalance became ever-more acute. Although Germany would revalue in October 1969 and Japan in July 1967, the pressures continued.

More academic economists began to echo the early calls by Friedman (1953) and Meade (1955) for floating exchange rates, arguing that market-determined rates would tend to eliminate external payments imbalances while insulating countries from foreign inflationary shocks. Their basic argument was that routine exchange-rate flexibility allows all countries to move to a preferred resolution of the trilemma—as compared with the situation of much more constrained policymaking that they then faced. As Johnson (1969, p. 18) put it: "Flexible rates would allow each country to pursue the mixture of unemployment and price trend objectives it prefers, consistent with international equilibrium, equilibrium being secured by appreciation of the currencies of 'price stability' countries relative to the currencies of 'full employment' countries."

By March 1973, after several attempts by the industrial countries to shore up fixed exchange rates, further co-operation proved impossible. Generalized floating

[5] Dagher (2016) discusses the political economy of deregulation following crises.

Figure 4
Fixed Exchange Rates and Capital Mobility, 1970–Present



Notes: Data from Shambaugh (2004) coding and Chinn and Ito (2006) database.

exchange rates emerged as a stopgap measure in the face of continuing speculative attacks. What was at the time intended as a temporary retreat has now lasted more than four decades.

Floating Exchange Rates: Monetary Independence and Financial Instability

The monetary trilemma implies that, with the imperative of exchange rate stability gone, countries in the 1970s could orient monetary policy toward domestic goals while still allowing additional freedom of capital movements across borders. In the decades since 1973, both exchange-rate flexibility and capital mobility have increased, but the process has not been smooth or consistent around the world. The United States financial account was already reasonably unrestricted at the start of the 1970s. European countries like Germany and Switzerland had imposed some inflow capital controls earlier, but could now dismantle them, whereas other European countries and Japan retained heavier controls through the late 1970s (Britain) or even up to the late 1980s (Bakker 1996; Abdelal 2007). As shown in Figure 4, the share of countries with pegged exchange rates fell dramatically from about 90 percent in 1970 to about 40 percent by the 1980s. But since then, the share of countries with pegged currencies has crept up over time to more than half. Conversely, the level of capital mobility was still relatively low in the mid-1980s, but then rose dramatically into the early 2000s, before leveling off and even declining during the last decade or so.

Although it was clearly feasible for countries to liberalize capital accounts once they had abandoned exchange rate pegs, it was not obvious that such a choice would be desirable, and outcomes have not been uniform. Many countries kept some form of pegged exchange rates, most of them emerging economies and developing countries, but also notably many European countries that established their own fixed exchange rate system in the 1970s, a precursor of the euro. In recent years, more countries have chosen to limit capital flows, notably after the 2008 global crisis. Volatility in exchange rates, in international capital flows, or in both can bring risks of financial and economic instability, as economic history has shown.

The Promise and Reality of Free International Capital Flows

In the 1970s, economists who made the case for capital account liberalization tended to stress the upside, emphasizing, for example, the negative effects of capital-control regimes that enabled governments to use financial repression to protect the domestic markets for their debts (McKinnon 1973; Shaw 1973). Moreover, capital controls became harder to enforce in the 1970s as domestic financial institutions developed and trade expanded further. The growing political clout of financial-sector interests also pushed in the direction of deregulation. More recently, Rajan and Zingales (2003) have suggested a narrative in which financial openness drives domestic liberalization by allowing greater competition in the financial sector and eroding the politically powerful interests that inhibit domestic reform to protect their rents.

However, the literature making the case for opening the capital account also often emphasized a desirable sequence of events, which began with liberalizing competition and establishing prudential regulation in domestic financial markets, and only then moving to openness to international capital flows. In their survey of the liberalization experience of 34 developing and advanced economies between 1973 and 1996, Williamson and Mahar (1998) found that, while most "liberalized the capital account gradually—after financial liberalization had occurred—in accord with the prevailing policy recommendation" (p. 31), "[f]ew countries seem to have heeded the advice to precede financial liberalization with the introduction of a system of prudential supervision, staffed by supervisors who have a high degree of independence of the political authorities" (p. 29). The piecemeal natures of some liberalizations contributed to later financial instability in some cases. Moreover, in the 1970s and 1980s, the accepted wisdom often did not emphasize interactions between opening the capital account and the need for considerable exchange rate flexibility.

The doctrinal shift regarding capital mobility seen in advanced economies in the 1970s and 1980s began to spread globally in the 1990s. By September 1997, the IMF's management was proposing that the Fund's executive board amend the Articles of Agreement to give the Fund an explicit role in guiding countries toward more open capital accounts. To be clear, the proposal was *not* advocating an indiscriminate rush toward opening; indeed, it recognized the role of capital inflows in financial crises, such as those that had afflicted Latin America from the mid-1970s through the mid-1990s, and it therefore explicitly sanctioned gradualism, based on

country circumstances (Fischer 1997). But it took as a given that an open capital account was the desirable ending point for all countries.

However, the 1997–98 financial crisis that rocked countries across East Asia marked an inflection point in economists' thinking about the merits of international capital mobility. With the Latin American debt crisis of the 1980s, one could make an argument that macroeconomic policymaking in those countries had been unsound, that their growth prospects had been overstated, and that they were prone to structural rigidities—a seemingly sufficient explanation for why their overborrowing came to grief. But the emerging economies of East Asia, without such apparent macroeconomic flaws, had seemed to provide shining examples of mostly well-run economies with rapid economic growth. These economies featured at least partially open capital accounts, which allowed for substantial inflows of foreign capital. They also had heavily managed exchange rates. These economies experienced what became known as a "sudden stop," when foreign (and often, domestic) capital fled these countries. The result was a drop in exchange rates, which made it impossible to repay dollar-denominated debt, triggering a meltdown of their financial markets and banking systems—financial dynamics for which there actually was ample precedent in the earlier Latin American crises (Díaz Alejandro 1985; Kaminsky and Reinhart 1999).

After the East Asian crisis, it became commonplace for economists (and the IMF) to recommend floating for such emerging and liberalizing economies (Fischer 2003). But in addition, the certitude that freeing the capital account should be a long-term goal for all countries fell by the wayside. Since then, there has been considerable rethinking of the doctrine as well as an accumulation of empirical evidence on capital account liberalization (for example, Ostry et al. 2010; Ocampo 2015). The International Monetary Fund (2012) published a new "institutional view" on capital controls, which sanctioned their use in some circumstances.

The Promise and Reality of Floating Exchange Rates

Early advocates of floating exchange rates like Friedman and Johnson clearly oversold the extent to which they could facilitate trade while still insulating a domestic economy from international shocks. They erred in part because, in their times, they had no immediate experience with the types of global financial shocks that have become more prevalent. Indeed, as shown earlier in Figure 4, a substantial number of countries have been unwilling to allow their currencies to float freely, and the prevalence of pegged currencies has exceeded half in the last decade or so. Presumably, those who peg their currencies believe that this choice will facilitate trade and protect their economy from macro-financial shocks caused by large exchange rate fluctuations, the essence of "fear of floating" (Calvo and Reinhart 2002).

Even early in the floating rate era, the new risks to financial stability were apparent. In June 1974, German regulators closed a small bank, the Bankhaus I. D. Herstatt, which had taken large foreign exchange positions far in excess of its capital. Later that year, the Franklin National Bank of New York also closed after foreign exchange losses. Interestingly, the Federal Reserve had to borrow from

European central banks to help Franklin National meet its obligations, a direction of funding that would be reversed when the Fed lent dollars to foreign central banks during the global financial crisis of the late 2000s.

But flexible exchange rates have their advantages, too. As noted a moment ago, the consequences of the "sudden stop" of capital inflows in the East Asian financial crises of 1997–98 was made worse because exchange rates had been heavily managed, and domestic banks and other financial institutions were unhedged and unprepared for a dramatic swing in exchange rates. In addition, as the monetary trilemma suggests, floating exchange rates empower domestically oriented monetary policy, while providing a shock absorber against external macroeconomic shocks.

For most countries around the world, one of the most potent external macroeconomic shocks involves changes in policy by the US Federal Reserve. Early work by Jay Shambaugh, and the three of us together, examined the empirical correlation between short-term and policy interest rates in home countries versus in "base" countries like the United States in modern times (Obstfeld, Shambaugh, and Taylor 2004, 2005; Shambaugh 2004). We looked at whether the bilateral exchange rate regime between the home and the base country in a given time period was a float or a peg, and whether the capital account was largely open or closed. In our panel data, for the home-base pairs and periods studied—covering advanced and emerging economies, and spanning epochs from the pre–World War I gold standard era to the post–Bretton Woods era of today—the clear result was that pegs with open capital accounts had much higher (and more statistically significant) interest rate correlations between them than did either floating exchange rates or pegs with closed capital accounts, which is consistent with what the monetary trilemma would predict. Other work on international transmission of interest rates has confirmed these findings, with a range of studies finding bigger responses of short-term interest rates for pegs versus floats.[6]

To what extent does the decoupling of short-term interest rates that floating allows carry over to macroeconomic outcomes? Probably the most important macroeconomic outcome variable is aggregate output, and di Giovanni and Shambaugh (2008) found evidence that when the home economy has an open capital account and a peg, it tends to experience a real GDP growth slowdown when its base country tightens monetary policy, whereas when the home country has a floating exchange rate or a peg with a closed capital account, such an effect is weak or nonexistent. This finding indicates a macroeconomic buffering role for floating exchange rates.

One recent branch of the research literature argues that the choice of exchange regime may not matter. Indeed, Rey (2013, 2016) suggests that the monetary trilemma may now have been transformed into a dilemma, writing that "*cross-border flows and leverage of global institutions transmit monetary conditions globally, even under floating exchange-rate regimes*" (Rey 2013, p. 310, emphasis in original).

[6]For example, see Borensztein, Zettelmeyer, and Philippon (2001); Frankel, Schmukler, and Servén (2004); Miniane and Rogers (2007); di Giovanni and Shambaugh (2008); Klein and Shambaugh (2015); Obstfeld (2015); Caceres, Carrière-Swallow, and Gruss (2016); Ricci and Shi (2016).

In this view, the key choice is between domestic control over monetary policy and openness to international capital flows, and the choice of exchange rate regime plays at most a secondary role. We agree that floating exchange rates do not offer a complete buffer against transmission of all international financial and monetary shocks. For example, Miranda-Agrippino and Rey (2015), Passari and Rey (2015), and Rey (2013, 2016) show that even in major, advanced, floating-rate economies there appears to be significant spillover from US interest rates, to the global financial cycle, to domestic macroeconomic, and to financial conditions. However, when faced with external shocks, countries with floating exchange rates still have a shock absorber that countries that peg exchange rates lack and thus can achieve preferred policy outcomes even if they cannot achieve full insulation of their economies (Obstfeld 2015). In this sense, more flexible exchange rates do provide a degree of differential insulation from external monetary shocks, as the monetary policy trilemma predicts. Adding further weight to this argument, Obstfeld, Ostry, and Qureshi (2017) document that in emerging markets, which are most vulnerable to external forces, global changes in risk sentiment have less effect on most domestic financial variables when the exchange rate regime is a free or managed float.

International Financial Stability and the Financial Trilemma

The classic monetary policy trilemma emphasizes that the combination of floating exchange rates and capital mobility will empower monetary policy to focus on domestic objectives. However, the monetary trilemma does not speak directly to financial stability concerns. Indeed, monetary policy alone may be a relatively ineffective tool for addressing potential financial stability problems. In this case, exposure to global financial shocks and cycles, perhaps the result of monetary or other developments in the industrial-country financial markets, may overwhelm countries even when their exchange rates are flexible. If this outcome is a risk, countries may desire some combination of financial regulations or restrictions on international capital mobility to shield their economies more fully.[7]

Concerns about the need for international coordination of bank regulation emerged almost immediately after the collapse of the Bretton Woods arrangements. The first meeting of the Basel Committee on Banking Supervision was held in February 1975. Since then, this group has worked to apportion regulatory authority among national supervisors to avoid gaps in oversight; to promote informational exchanges; and to regularize international best practice in regulation, including standards for capital. There have been three successive initiatives on bank capital and other regulations starting in 1988. The Basel Committee has expanded over time and has drawn emerging markets into its orbit. Supplementing the work of the Basel Committee, and housed along with it at the Bank for International Settlements, is the Financial Stability Board, which originated in 1999 as the Financial Stability Forum and monitors the broader international financial system. The

[7]This is the core argument that Rey (2013, 2016) makes; and on global financial cycles, see also Borio and Disyatat (2015), Avdjiev, McCauley, and Shin (2016), and Reinhart, Reinhart, and Trebesch (2016).

work of the Financial Stability Board has become ever more important as "shadow banking" has grown up alongside more traditional banking, as the range of financially systemic and globally active institutions has expanded, and as the complexity of financial markets and the instruments traded in them has grown.

The growing efforts of international regulators to coordinate on financial oversight have been mirrored in the rise to prominence of the *financial trilemma* (Schoenmaker 2013), which is distinct from the monetary policy trilemma discussed above. In the financial trilemma, countries must choose between national financial policies, integration into global financial markets, or financial stability. For example, if there is widespread integration into global financial markets and each nation retains national sovereignty over financial policies, then regulatory arbitrage among jurisdictions may undermine financial stability (for some evidence on arbitrage channels, see Aiyar, Calomiris, and Wieladek 2014; Bayoumi 2017; Cerutti, Claessens, and Laeven 2017). Alternatively, a country with national financial rules may enhance financial stability by cutting off integration into global markets. However, most countries have been willing effectively to surrender a certain amount of sovereignty over financial regulation in the hope of keeping access to international capital markets while maintaining financial stability. While the financial trilemma obviously applies to currency zones with integrated payments systems like the euro area, it also applies to countries that maintain their own floating currencies.

Because of the *financial* trilemma, moreover, domestic monetary policy, under an open capital account and a floating exchange rate, even if more autonomous than under a pegged exchange rate, will likely face a harsher tradeoff between conventional macroeconomic goals (inflation, output) and financial stability (Obstfeld 2015). Thus, the burden on domestic financial stability policy will accordingly be even greater. Macroprudential policies must bear some of the load, and in the face of certain kinds of shocks, some forms of capital controls could appear desirable as well, as argued, for example, by Blanchard (2016).

The European Example: Pegged or Stable Exchange Rates and Financial Fragility

The nations of western Europe have charted a hybrid path for monetary institutions in the post–Bretton Woods era. After the early 1970s breakdown of fixed rates, the members of what was at the time called the European Economic Community (EEC) moved to limit currency fluctuations within their group. Indeed, as early as 1969 these countries were already contemplating the Werner Plan for ultimately moving to a common currency. By 1979, an important subset of EEC members pegged their mutual exchange rates in what was known as the European Monetary System. The resulting exchange rate mechanism ended up functioning much like a miniature Bretton Woods system—with periodic crises and exchange rate parity adjustments, only now with Germany as the center country. In line with the trilemma, some members, including France and Italy, maintained capital controls.

By the end of the 1980s, as was later codified in the Maastricht Treaty of 1991, momentum built for the move toward so-called *economic* and *monetary union*. Concretely, the former meant a single market concept under which capital controls

had to disappear; the latter meant that fixed exchange rates became the overriding objective of national monetary policies, as a stepping stone to the common currency. The future members of the euro area thus embraced the vertex of the monetary policy trilemma based on capital mobility and exchange-rate stability vis-à-vis each other, but with jointly floating exchange rates against outside currencies. Abdelal (2007) offers an insightful treatment of the European attitude toward capital controls, and its impact on global practice more generally.

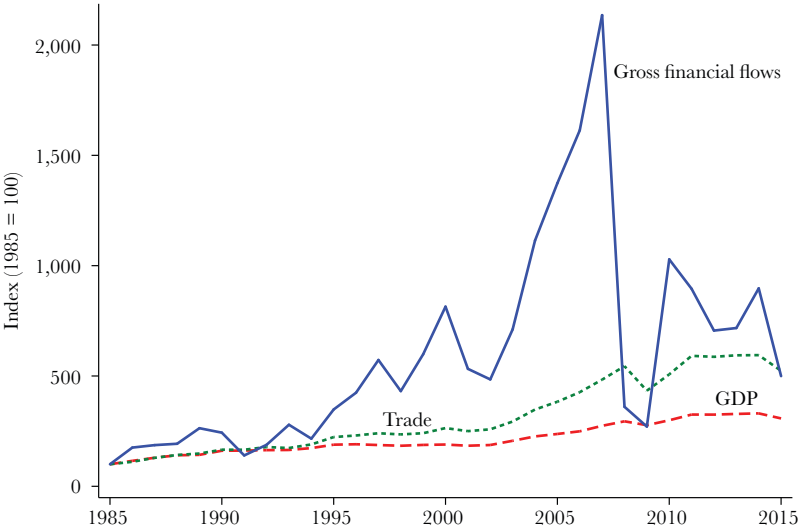
However, just as the earlier Bretton Woods treaty had neglected financial stability concerns, the Maastricht Treaty of 1991 setting up the European economic and monetary union likewise turned a blind eye to financial stability (as opposed to macroeconomic) issues, in a different setting and for different reasons, but with considerable destabilizing effects later. There was no mechanism built into the euro to address a situation in which some countries ran continuous and large trade surpluses while others ran large and continuous deficits. There was no common framework of prudential banking and financial regulation, much less any pooling of bank failure risk (for example, deposit insurance). And, as became evident in subsequent euro area crises, banks and governments could even run out of liquidity despite the single currency, amplifying financial stability risks. Unlike the 1950s and 1960s, when a quite repressive global financial environment ensured that the neglect of these issues under Bretton Woods would not prove too costly, the disregard for financial stability in the euro architecture in a time of rampant financialization would prove to be a painful oversight.

Old Problems in New Guises

The causes of the global financial crisis of 2007–2008 have been much debated. The financial boom that preceded the crash of 2007–2008 was a global phenomenon. Bernanke (2005) argued that the world economy was experiencing a global saving glut, driven primarily by China and the former crisis countries of East Asia, distributing ample liquidity worldwide and pushing up real estate prices in many countries, not just the United States. But this emphasis on *net* capital flows from countries with surpluses of saving over investment obscured another prominent feature of the period, the sharp rise in *gross* capital (largely bank-related) flows between countries that helped to prepare the ground for the subsequent crash (Bernanke, Bertaut, DeMarco, and Kamin 2011; Lane 2012; Borio, James, and Shin 2014). Figure 5 illustrates the behavior of these flows leading up to the financial crisis and after.

Despite having exchange-rate flexibility as a potential brake, some countries were unable to head off the resulting amplification of financial instability coming through open capital markets. Within the euro area, with no exchange rates at all to adjust, cross-border capital flows from core to periphery played a major destabilizing role, notably in the credit booms of Ireland and Spain (Lane 2013; Hale and Obstfeld 2016). Moreover, as advanced economies turned to ultra-loose monetary

Figure 5
Evolution of Real Gross Capital Flows Compared with Output and Trade, 1985–2015



Source: IMF World Economic Outlook and International Financial Statistics databases.
Notes: Indices are calculated from data in real US dollars (deflated using US GDP deflator). Global trade is defined as the average of global exports and imports of goods and services. Gross global financial flows are defined as the sum of direct investment, portfolio investment, and other investments. Values are obtained by averaging inflows and outflows to account for measurement error.

policies in the wake of the financial crisis, some emerging markets, while having loosened the rigidity of their exchange rates after the Asian crisis, still found that lower global interest rates and capital inflows were making it harder for them to maintain financial and price stability. The central macroeconomic challenges of exchange rate regime choice, external payments adjustment, and international liquidity have clearly remained over time, although they have manifested themselves in different forms given the evolution of financial markets.

How Should Exchange Rates Be Determined?

A number of countries have continued to use some form of pegged exchange rates, as shown earlier in Figure 4. However, the monetary trilemma, coupled with widespread financial integration, has made it much harder—or even impossible, for most countries—to maintain completely firm currency pegs, given the imperatives of domestically oriented monetary policy. At the national level, as we have seen, floating exchange rates clearly cannot provide insulation against all global financial or real shocks. But floating still does facilitate some measure of domestic insulation, and policymakers can provide additional shock absorbers by deploying effective financial and macroprudential policies, by adopting sound fiscal and structural policies, and even by using measures to limit capital flow in some circumstances.

But while floating or soft peg exchange rates have helped mitigate policymakers' domestic challenges, debate has continued over whether floating is a suitable solution for the international system as a whole. While floating exchange rates can allow individual countries to stabilize to a degree, they also raise the age-old problem of competitive currency depreciations, in which demand is just being shifted between countries. Central bankers faced with this "currency war" critique also typically respond that while monetary expansion and lower interest rates within a country indeed do depreciate the domestic currency and make foreign goods relatively more expensive, the lower domestic interest rates also bring about a win-win rise in domestic demand (via the interest rate channel) that spills over positively abroad. This argument may appear to lose traction in today's economy where major central bank policy interest rates have settled near their effective lower bounds.[8] But the arrival at that unpleasant floor is a result of other factors, notably the conjunction of low real rates and current inflation targets, and not a mark against conventional policy in normal times per se.

Low global real interest rates, however, reflect the balance between global saving and global investment, and for each individual country, its current account surplus equals the excess of its saving over its investment. These facts raise the concern that some economies may be boosting their economies through higher trade surpluses, pushing global real interest rates down and making monetary stabilization more difficult for all.

How Should Balance of Payments Adjustments Occur?

Countries with large trade deficits, experiencing an inflow of foreign investment capital, face the threat of "sudden stop," and therefore have some incentives to limit their external imbalance. On the other hand, there is no such market-based incentive to limit trade surpluses. In a world where high balance-of-payments surpluses persist for certain countries, net external wealth positions become increasingly divergent. Creditors' external wealth becomes ever more positive, and debtors' becomes ever more negative, with debtor efforts to fend off deflation only prolonging the process. When economies that have been experiencing large and sustained current account deficits eventually are forced to adjust spending abruptly when their perceived intertemporal budget constraints shift, as is often the case, the result can be national or even international recession and crisis.

In some cases, one country's higher trade surplus may come directly at the expense of employment and price stability abroad (Caballero, Farhi, and Gourinchas 2015; Eggertsson, Mehrotra, Singh, and Summers 2016). The problem is less serious when countries can deploy monetary or other policies to offset deflationary impulses from abroad (Blanchard and Milesi-Ferretti 2012). For an economy at the

[8] Mishra and Rajan (2016) suggest that the unconventional monetary policies employed at the effective lower bound for monetary policy interest rates may in some cases work primarily by shifting aggregate demand from other countries, rather than stimulating interest-sensitive expenditure components at home, and that policies which are globally zero sum should be avoided.

effective lower bound of the policy interest rate, however, monetary policy alone cannot easily offset a foreign deflationary impulse; moreover, a fiscal policy response may be constrained as well by fears (justified or not) over pre-existing high levels of public debt.

The problems that arise when some countries run sustained and large trade imbalances have been well-understood by economists since at least the interwar years, but this issue has repeatedly proven intractable to global macroeconomic policy solutions. International cooperation is at a much more evolved stage with respect to trade policy (the World Trade Organization with its rules and oversight) and in financial regulatory policy (the Basel process and the Financial Stability Board), and is even advancing in international tax policy. One reason may be that the gains from those other modes of cooperation potentially accrue simultaneously to all parties. However, the identities of countries with large trade surpluses tend to be fairly persistent over time, giving them less incentive to submit to rules or suasion today in the expectation that someday they may be running deficits. Another reason may be that economists and policymakers have, rightly or wrongly, more precise expectations about the nature and effects of trade, regulatory, and tax instruments, compared with macroeconomic policy tools.

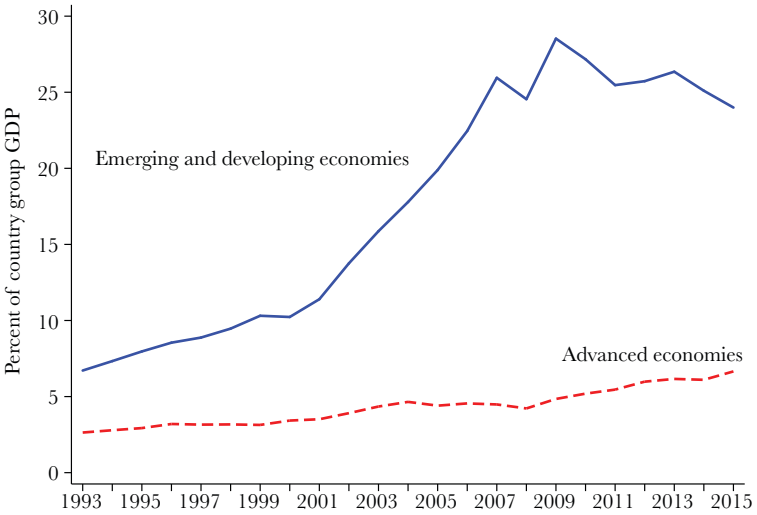
How Can Countries Have Access to Adequate International Liquidity?

Under the Bretton Woods system, countries held foreign exchange reserves (mostly in US dollar assets) to peg their exchange rates. Accordingly, the advent of floating exchange rates led many economists to predict that central banks would reduce their demands for reserves. As in the last few decades, aside from the role that reserves play in foreign exchange intervention, they can also play a potential role in buffering balance of payments shocks when other means of external financing become expensive or unavailable, for example, in a sudden stop. Here as well, the development of international capital markets after the early 1970s led some to predict that expanded opportunities for foreign borrowing would reduce the role of reserves.

Such predictions have been wildly wrong, as we can see from Figure 6. Advanced-country reserves remain significant relative to GDP, rising from about 3 percent of GDP in 1993 for these countries as a group, to more than 5 percent of GDP by 2015. However, the reserves held by emerging and developing countries have risen sharply, rising from about 7 percent of the GDP of this group of countries in 1993 to about 25 percent of their GDP by 2007—and remaining roughly at that level since then (based on IMF data).

Emerging and developing economies have raised their reserve holdings for two main reasons. First, even though their exchange rates have generally become more flexible in the last few decades, they continue to intervene in foreign exchange markets. In some cases, the goal has been to temper currency volatility (Calvo and Reinhart 2002); in others, to maintain or enhance export competitiveness—motivations that in practice can overlap. Second, more open international capital markets have *raised* the precautionary demand for reserves, not reduced it. For emerging

Figure 6
Stocks of International Reserves, 1993–2015



Sources: IMF International Financial Statistics database for reserve data (which include gold valued using national methods); IMF World Economic Outlook database for GDP data.
Note: The "advanced" group excludes Hong Kong, South Korea, Singapore, and Taiwan but includes the Czech Republic, Estonia, Slovenia, and the Slovak Republic.

market economies, larger balance-sheet liabilities, some denominated in foreign currencies and at short term, imply a greater risk of capital-flow reversal: not only might financing for a current account deficit disappear in a sudden stop, but foreign creditors could also call for the repayment of gross liabilities. In addition, domestic investors might seek to rebalance portfolios towards foreign assets, via capital flight towards perceived safe havens. The magnitudes of these gross flows can greatly exceed those of net flows, and these risks increase the utility of foreign exchange reserves to help domestic financial institutions as well as importers make payments abroad, while minimizing the risks of possible spillovers to domestic banks (Obstfeld, Shambaugh, and Taylor 2010).

Such risks are not limited to emerging and developing economies. Banks worldwide fund themselves with borrowing in key advanced-economy currencies, notably the US dollar, which continues to play a pivotal international role long after the Bretton Woods system's demise.[9] During the global financial crisis, for example, European banks found it difficult to roll over short-term US dollar credits, and faced the prospect of having to liquidate dollar-denominated assets in fire sale conditions. Ad hoc swap lines, through which the Federal Reserve lent dollars, and with which

[9] A prescient meditation on the centrality of the US dollar, still relevant 50 years later, is Kindleberger (1967).

foreign central banks could meet these needs (and assume the attached credit risk), helped stabilize markets. Indeed, these arrangements became permanent late in 2013 among the six key advanced-economy central banks. Helpful and necessary as this arrangement is, it still leaves emerging-market central banks out in the cold (Weder di Mauro and Zettelmeyer 2017).

The existing system of gross reserve holding by emerging-market central banks has several drawbacks, discussed in detail in Obstfeld (2013), among which is the risk that large-scale reserve accumulation is deflationary globally. These problems could be ameliorated if instead emerging and developing countries had better access to credit lines. Traditional IMF lending cannot fulfill this role, as IMF programs are subject to conditionality and time-consuming negotiation. Over the years, the IMF has tried to offer various more-flexible credit facilities for prequalified borrowers, but few countries have signed up, fearing either the stigma of asking for a credit line or of receiving one and later being disqualified. In any case, a globally systemic crisis would strain the Fund's capacity. The desire of nonadvanced economies to hold higher reserves raises a modern-day analog of the Triffin paradox from the 1960s (Farhi, Gourinchas, and Rey 2011; Obstfeld 2013). Reserves these days mostly take the form of high-quality "safe" liabilities of advanced countries, generally government-issued or -guaranteed. But the supply of these liabilities is not unlimited; indeed, it has arguably shrunk as several advanced-country governments, notably in the euro area, became fiscally challenged after the crises of 2008–2012. Just as the Triffin dilemma during the 1960s was that the United States could not continue to satisfy the world's growing demand for dollar reserves without undermining its commitment to convert them into gold, so the advanced-economy reserve issuers cannot issue unlimited amounts of reserve claims without undermining the "safe asset" character of those liabilities that makes them useful as reserve assets in the first place. There is little doubt that excess global demand for safe assets, including safe reserve assets, is contributing to the current low interest rate environment in the world economy.

Summing Up

One of the most important realizations to come out of the global financial crisis of 2007–09 and its aftermath was that standard models of macroeconomic stabilization had not paid sufficient attention to finance and financial markets. A similar realization holds for models of international monetary relations. In both cases, policy practice and intellectual debate have been struggling for centuries to address financial stability concerns. In the last few decades, the task has become even more urgent in the face of rapidly evolving financial markets, seemingly intent on pushing risky activities outside the perimeters of regulation. Economic analysis still needs to bring the risks of financial instability into its core frameworks, from the analysis of business cycles to that of international economic interactions.

■ *This paper reflects the views of its authors alone, and not those of the IMF, its management, or its executive board. For their helpful comments and criticism, we thank Ben Bernanke, Olivier Blanchard, Claudio Borio, Jihad Dagher, Stanley Fischer, Patrick Honohan, Michael Klein, José Antonio Ocampo, Carmen Reinhart, and Dirk Schoenmaker. For discussion and assistance, we thank Helge Berger, Eugenio Cerutti, Chanpheng Fizzarotti, Jonathan Ostry, Hui Tong, and Haonan Zhou. Mark Gertler, Gordon Hanson, and Timothy Taylor offered expert editorial guidance. All errors are ours.*

References

Abdelal, Rawi. 2007. *Capital Rules: The Construction of Global Finance.* Harvard University Press.

Aiyar, Shekhar, Charles W. Calomiris, and Tomasz Wieladek. 2014. "Does Macro-Prudential Regulation Leak? Evidence from a UK Policy Experiment." *Journal of Money, Credit, and Banking* 46(S1): 181–214.

Avdjiev, Stefan, Robert N. McCauley, and Hyun Song Shin. 2016. "Breaking Free of the Triple Coincidence in International Finance." *Economic Policy* 31(87): 409–51.

Bagehot, Walter. 1873. *Lombard Street: A Description of the Money Market.* Henry S. King and Co.

Bakker, Age F. P. 1996. *The Liberalization of Capital Movements in Europe: The Monetary Committee and Financial Integration, 1958–1994.* Kluwer Academic Publishers.

Bayoumi, Tamim. Forthcoming. *The Unexplored Causes of the Financial Crisis and the Lessons Yet to Be Learned.* Yale University Press.

Bernanke, Ben S. 1983. "Nonmonetary Effects of the Financial Crisis in Propagation of the Great Depression." *American Economic Review* 73(3): 257–76.

Bernanke, Ben S. 2005. "The Global Saving Glut and the U.S. Current Account Deficit." Lecture presented at the Sandridge Lecture, Virginia Association of Economists, Richmond, Virginia, March 10.

Bernanke, Ben S., Carol Bertaut, Laurie Pounder DeMarco, and Steven Kamin. 2011. "International Capital Flows and the Returns to Safe Assets in the United States, 2003–2007." *Revue de la Stabilité Financière* 15: 15–30.

Bernanke, Ben S., and Kevin Carey. 1996. "Nominal Wage Stickiness and Aggregate Supply in the Great Depression." *Quarterly Journal of Economics* 111(3): 853–83.

Bernanke, Ben S., and Harold James. 1991. "The Gold Standard, Deflation, and Financial Crisis in the Great Depression: An International Comparison." In *Financial Markets and Financial Crises,* edited by R. Glenn Hubbard, 33–68. University of Chicago Press.

Blanchard, Olivier. 2016. "Currency Wars, Coordination, and Capital Controls." NBER Working Paper 22388.

Blanchard, Olivier, and Gian Maria Milesi-Ferretti. 2012. "(Why) Should Current Account Balances Be Reduced?" *IMF Economic Review* 60(1): 139–50.

Bordo, Michael D., and Anna J. Schwartz, ed. 1984. *A Retrospective on the Classical Gold Standard, 1821–1931.* University of Chicago Press.

Borensztein, Eduardo R., Jeromin Zettelmeyer, and Thomas Philippon. 2001. "Monetary Independence in Emerging Markets: Does the Exchange Rate Regime Make a Difference?" IMF Working Paper WP/01/1.

Borio, Claudio, and Piti Disyatat. 2015. "Capital Flows and the Current Account: Taking Financing (More) Seriously." BIS Working Paper 525.

Borio, Claudio, Harold James, and Hyun Song Shin. 2014. "The International Monetary and Financial System: A Capital Account Historical Perspective." BIS Working Paper 457.

Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2015. "Global Imbalances and Currency Wars at the ZLB." NBER Working Papers 21670.

Caceres, Carlos, Yan Carrière-Swallow, and Bertrand Gruss. 2016. "Global Financial Conditions and Monetary Policy Autonomy." IMF Working Paper WP/16/108.

Calvo, Guillermo A., and Carmen M. Reinhart. 2002. "Fear of Floating." *Quarterly Journal of Economics* 117(2): 379–408.

Campa, Jose Manuel. 1990. "Exchange Rates and Economic Recovery in the 1930s: An Extension to Latin America." *Journal of Economic History* 50(3): 677–82.

Cassis, Youssef. 2011. *Crises and Opportunities: The Shaping of Modern Finance.* Oxford University Press.

Cerutti, Eugenio, Stijn Claessens, and Luc Laeven. 2017. "The Use and Effectiveness of Macroprudential Policies: New Evidence." *Journal of Financial Stability* 28: 203–24.

Chinn, Menzie D., and Hiro Ito. 2006. "What Matters for Financial Development? Capital Controls, Institutions, and Interactions." *Journal of Development Economics* 81(1): 163–92.

Cooper, Richard N. 1982. "The Gold Standard: Historical Facts and Future Prospects." *Brookings Papers on Economic Activity* 1: 1–45.

Dagher, Jihad C. 2016. "Regulatory Cycles: Revisiting the Political Economy of Financial Crises." Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2772373.

di Giovanni, Julian, and Jay C. Shambaugh. 2008. "The Impact of Foreign Interest Rates on the Economy: The Role of the Exchange Rate Regime." *Journal of International Economics* 74(2): 341–61.

Díaz-Alejandro, Carlos F. 1985. "Good-bye Financial Repression, Hello Financial Crash." *Journal of Development Economics* 19(1–2): 1–24.

Eggertsson, Gauti B., Neil R. Mehrotra, Sanjay R. Singh, and Lawrence H. Summers. 2016. "A Contagious Malady? Open Economy Dimensions of Secular Stagnation." *IMF Economic Review* 64(4): 581–634.

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

Eichengreen, Barry. 2008. *Globalizing Capital: A History of the International Monetary System,* 2nd edition. Princeton University Press.

Eichengreen, Barry, and Jeffrey Sachs. 1985. "Exchange Rates and Economic Recovery in the 1930s." *Journal of Economic History* 45(4): 925–46.

Farhi, Emmanuel, Pierre-Olivier Gourinchas, and Hélène Rey. 2011. *Reforming the International Monetary System.* Centre for Economic Policy Research.

Fischer, Stanley. 1997. "Capital Account Liberalization and the Role of the IMF." Paper presented at the Asia and the IMF Seminar, Hong Kong, China, September 19.

Fischer, Stanley. 2003. "Financial Crises and Reform of the International Financial System." *Review of World Economics/Weltwirtschaftliches Archiv* 139(1): 1–37.

Frankel, Jeffrey, Sergio L. Schmukler, and Luis Servén. 2004. "Global Transmission of Interest Rates: Monetary Independence and Currency Regime." *Journal of International Money and Finance* 23(5): 701–33.

Friedman, Milton. 1953. "The Case for Flexible Exchange Rates." In *Essays in Positive Economics*, 157–203. University of Chicago Press.

Friedman, Milton, and Anna Jacobson Schwartz. 1963. *A Monetary History of the United States, 1867–1960.* Princeton University Press.

Hale, Galina, and Maurice Obstfeld. 2016. "The Euro and the Geography of International Debt Flows." *Journal of the European Economic Association* 14(1): 115–44.

Hamilton, James D. 1988. "Role of the International Gold Standard in Propagating the Great Depression." *Contemporary Policy Issues* 6(2): 67–89.

Hetzel, Robert L., and Gary Richardson. 2016. "Money, Banking, and Monetary Policy from the Formation of the Federal Reserve until Today." Federal Reserve Bank of Richmond Working Paper 16–1.

Hsieh, Chang-Tai, and Christina D. Romer. 2006. "Was the Federal Reserve Constrained by the Gold Standard during the Great Depression? Evidence from the 1932 Open Market Purchase Program." *Journal of Economic History* 66(1): 140–76.

International Monetary Fund. 2012. *The Liberalization and Management of Capital Flows: An Institutional View.* Washington, DC: International Monetary Fund.

Irwin, Douglas A. 2011. *Trade Policy Disaster: Lessons from the 1930s.* MIT Press.

Irwin, Douglas A. 2012–2013. "The French Gold Sink and the Great Deflation of 1929–32." *Cato Papers on Public Policy* 2: 1–56.

Jalil, Andrew J. 2015. "A New History of Banking Panics in the United States, 1825–1929: Construction and Implications." *American Economic Journal: Macroeconomics* 7(3): 295–330.

Johnson, Harry G. 1969. "The Case for Flexible Exchange Rates, 1969." *Federal Reserve Bank of St. Louis Review* 51(6): 12–24.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. Forthcoming. "Macrofinancial History and the New Business Cycle Facts." In *NBER Macroeconomics Annual 2016, Volume 31,* edited by Martin Eichenbaum and Jonathan A. Parker. University of Chicago Press.

Kindleberger, Charles P. 1967. "The Politics of International Money and World Language." *Essays in International Finance* 61: 1–16.

Kindleberger, Charles P. 1973 [2013]. *The World*

in Depression: 1929–1939. 40th Anniversary edition. University of California Press.

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

Keynes, John Maynard. 1925. *The Economic Consequences of Mr. Churchill*. Leonard and Virginia Woolf at the Hogarth Press.

Keynes, John Maynard. 1930 [2012]. *A Treatise on Money*, Vol. 2: *The Applied Theory of Money*. Vol. 6 of The Collected Writings of John Maynard Keynes. Cambridge University Press.

Klein, Michael W., and Jay C. Shambaugh. 2015. "Rounding the Corners of the Policy Trilemma: Sources of Monetary Policy Autonomy." *American Economic Journal: Macroeconomics* 7(4): 33–66.

Laidler, David. 2003. "Two Views of the Lender of Last Resort: Thornton and Bagehot." *Cahiers d'Economie Politique* 45: 61–78.

Lane, Philip R. 2012. "Financial Globalisation and the Crisis." BIS Working Paper 397.

Lane, Philip R. 2013. "Capital Flows in the Euro Area." European Commission European Economy Economic Paper 497.

League of Nations. 1944. *International Currency Experience: Lessons of the Interwar Period*. Princeton University Press.

McKinnon, Ronald I. 1973. *Money and Capital in Economic Development*. Brookings Institution Press.

Meade, James E. 1955. "The Case for Variable Exchange Rates." *Three Banks Review* 27: 3–27.

Miniane, Jacques, and John H. Rogers. 2007. "Capital Controls and the International Transmission of U.S. Money Shocks." *Journal of Money, Credit, and Banking* 39(5): 1003–35.

Miranda-Agrippino, Silvia, and Hélène Rey. 2015. "World Asset Markets and the Global Financial Cycle." NBER Working Paper 21722.

Mishkin, Frederic S. 1978. "The Household Balance Sheet and the Great Depression." *Journal of Economic History* 38(4): 918–37.

Mishra, Prachi, and Raghuram Rajan. 2016. "Rules of the Monetary Game." Department of Economic and Policy Research Reserve Bank of India RBI Working Paper 04/2016.

Obstfeld, Maurice. 2013. "The International Monetary System: Living with Asymmetry." In *Globalization in an Age of Crisis: Multilateral Economic Cooperation in the Twenty-First Century*, edited by Robert C. Feenstra and Alan M. Taylor, 301–36. University of Chicago Press.

Obstfeld, Maurice. 2015. "Trilemmas and Tradeoffs: Living with Financial Globalization." In *Global Liquidity, Spillovers to Emerging Markets and Policy Responses*, edited by Claudio Raddatz, Diego Saravia, and Jaume Ventura, 13–78. Central Bank of Chile.

Obstfeld, Maurice, Jonathan D. Ostry, and Mahvash S. Qureshi. 2017. "A Tie that Binds: Revisiting the Trilemma in Emerging Market Economies." IMF Working Paper WP/17/130, June.

Obstfeld, Maurice, Jay C. Shambaugh, and Alan M. Taylor. 2004. "Monetary Sovereignty, Exchange Rates, and Capital Controls: The Trilemma in the Interwar Period." *IMF Staff Papers* 51(Special Issue): 75–108.

Obstfeld, Maurice, Jay C. Shambaugh, and Alan M. Taylor. 2005. "The Trilemma in History: Tradeoffs among Exchange Rates, Monetary Policies, and Capital Mobility." *Review of Economics and Statistics* 87(3): 423–38.

Obstfeld, Maurice, Jay C. Shambaugh, and Alan M. Taylor. 2010. "Financial Stability, the Trilemma, and International Reserves." *American Economic Journal: Macroeconomics* 2(2): 57–94.

Obstfeld, Maurice, and Alan M. Taylor. 1998. "The Great Depression as a Watershed: International Capital Mobility over the Long Run." In *The Defining Moment: The Great Depression and the American Economy in the Twentieth Century*, edited by Michael D. Bordo, Claudia Goldin, and Eugene N. White, 353–402. University of Chicago Press.

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

Ocampo, José Antonio. 2015. "Capital Account Liberalization and Management." WIDER Working Paper 2015/048.

Ostry, Jonathan D., Atish Ghosh, Karl Habermeier, Marcos Chamon, Mahvash S. Qureshi, and Dennis B. S. Reinhardt. 2010. "Capital Inflows: The Role of Controls." International Monetary Fund Staff Position Note 10/04.

Padoa-Schioppa, Tommaso. 1988. "The European Monetary System: A Long-Term View." In *The European Monetary System*, edited by Francesco Giavazzi, Stefano Micossi, and Marcus Miller, 369–384. Cambridge University Press.

Passari, Evgenia, and Hélène Rey. 2015. "Financial Flows and the International Monetary System." *Economic Journal* 125(584): 675–98.

Polanyi, Karl. 1944. *The Great Transformation*. Farrar & Rinehart.

Qian, Rong, Carmen M. Reinhart, and Kenneth S. Rogoff. 2011. "On Graduation from Default, Inflation and Banking Crises: Elusive or Illusion?" In *NBER Macroeconomics Annual 2010,* vol. 25, edited by Daron Acemoglu and Michael Woodford, 1–36. University of Chicago Press.

Quinn, Dennis, Martin Schindler, and A. Maria Toyoda. 2011. "Assessing Measures of Financial Openness and Integration." *IMF Economic Review* 59(3): 488–522.

Rajan, Raghuram G., and Luigi Zingales. 2003. "The Great Reversals: The Politics of Financial Development in the Twentieth Century." *Journal of Financial Economics* 69(1): 5–50.

Reinhart, Carmen M., Vincent Reinhart, and Christoph Trebesch. 2016. "Global Cycles: Capital Flows, Commodities, and Sovereign Defaults, 1815–2015." *American Economic Review* 106(5): 574–80.

Rey, Hélène. 2013. "Dilemma Not Trilemma: The Global Financial Cycle and Monetary Policy Independence." Paper presented at the Global Dimensions of Unconventional Monetary Policy Federal Reserve Bank of Kansas City Symposium, Jackson Hole, WY, August 21–23.

Rey, Hélène. 2016. "International Channels of Transmission of Monetary Policy and the Mundellian Trilemma." *IMF Economic Review* 64(1): 6–35.

Ricci, Luca Antonio, and Wei Shi. 2016. "Trilemma or Dilemma; Inspecting the Heterogeneous Response of Local Currency Interest Rates to Foreign Rates." IMF Working Paper WP/16/75.

Ruggie, John Gerard. 1982. "International Regimes, Transactions, and Change: Embedded Liberalism in the Postwar Economic Order." *International Organization* 36(2): 379–415.

Schoenmaker, Dirk. 2013. *Governance of International Banking: The Financial Trilemma*. Oxford University Press.

Shambaugh, Jay C. 2004. "The Effect of Fixed Exchange Rates on Monetary Policy." *Quarterly Journal of Economics* 119(1): 301–52.

Shaw, Edward S. 1973. *Financial Deepening in Economic Development*. Oxford University Press.

Smith, Adam. 1776. *The Wealth of Nations*.

Temin, Peter. 1989. *Lessons from the Great Depression*. MIT Press.

Thornton, Henry. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. J. Hatchard.

Triffin, Robert. 1960. *Gold and the Dollar Crisis: The Future of Convertibility*. Yale University Press.

Weder di Mauro, Beatrice, and Jeromin Zettelmeyer. 2017. *CIGI Essays on International Finance: The New Global Financial Safety Net: Struggling for Coherent Governance in a Multipolar System*, vol. 4. Centre for International Governance Innovation.

Williamson, John, and Molly Mahar. 1998. "A Survey of Financial Liberalization." *Essays in International Finance* 211: 1–74.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 29–46

The Safe Assets Shortage Conundrum

Ricardo J. Caballero, Emmanuel Farhi, and Pierre-Olivier Gourinchas

Economic actors need stores of value. Households save for retirement, for a rainy day, or to transmit wealth to their offspring. Corporations need to hold cash. Financial institutions need collateral. Central banks and sovereign wealth funds need to hold foreign assets. These stores of value come in many forms: cash, bank deposits, US government Treasury bills, and also corporate bonds, stocks, repurchase agreements, derivatives, or real assets such as real estate, land, gold, and others.

All stores of value are not created equal. They differ in their degree of liquidity—the ease with which they can be traded—and in their sensitivity to various risk factors. Among the menu of available assets, some are perceived as "safer" than others. Yet safety is an elusive concept, because nothing is ever absolutely safe. Investors will always view the safety of an asset through the prism of their own perceptions, needs, and concerns, in relation to other assets, and in relation to the perceptions of other investors.

This paper adopts a pragmatic and narrow definition: a safe asset is a simple debt instrument that is expected to preserve its value during adverse systemic events (for example, Caballero and Farhi 2017). This operational definition captures the

■ *Ricardo J. Caballero is Ford International Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Emmanuel Farhi is Professor of Economics, Harvard University, Cambridge, Massachusetts. Pierre-Olivier Gourinchas is the S.K. and Angela Chan Professor of Economics, University of California, Berkeley, California. All three authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are caball@mit.edu, efarhi@fas.harvard.edu, and pog@berkeley.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.29

doi=10.1257/jep.31.3.29

emphasis of Dang, Gorton, and Holmström (2015) and Gorton (2016) on "information insensitivity," in the sense that safe assets can be transacted without much analysis or concern for adverse selection. It is also consistent with Caballero and Simsek's (2013) view that "simple" assets have special value during economic crises that are inherently complex. Finally, it also captures an essential strategic complementarity: an asset is safe if others expect it to be safe (Farhi and Maggiori 2016; He, Krishnamurthy, and Milbradt 2016). When it comes to forming beliefs about which assets are safe, reputations and history matter.

In modern economies, the financial sector and the government are the main manufacturers of financial assets: central banks issue cash and central bank reserves; Treasury departments issue government bonds and notes; banks and shadow banks issue short-term deposits or more complex instruments. The capacity of a country to produce safe assets is determined by constraints in the financial sector, the level of financial (under-) development, the fiscal capacity of the sovereign, and the track record of the central bank for exchange rate and price stability. For these reasons, the supply of safe assets, private and public, has historically been concentrated in a small number of advanced economies, most prominently the United States.

For the last few decades, with minor cyclical interruptions, the supply of safe assets has not kept up with global demand. The reason is straightforward: the collective growth rate of the advanced economies that produce safe assets has been lower than the world's growth rate, which has been driven disproportionately by the high growth rate of high-saving emerging economies such as China. If demand for safe assets is proportional to global output, this shortage of safe assets is here to stay.

The signature of this growing shortage is a steady increase in the price of safe assets, necessary to restore equilibrium in this market. Equivalently, global safe interest rates must decline, as has been the case since the 1980s. Simultaneously, we observed a surge in cross-border purchases of safe assets by safe asset demanders—many of them located in emerging economies—from safe asset producers, mostly the United States.

The early literature, brought to light by then–Federal Reserve vice-chair Bernanke's famous "savings glut" speech (Bernanke 2005), focused on a *general* shortage of assets without isolating its safe asset component (Caballero 2006). This literature aimed to explain the downward trend in interest rates as well as increasing global imbalances, that is the large current account deficits of the US economy and surpluses of Asian emerging markets. In Caballero, Farhi, and Gourinchas (2008), we showed how the endemic problem of a general shortage of assets in emerging markets was beginning to spread to the world at large through the large current account surpluses in Asian emerging markets. It was well understood then that a large share of these imbalances was caused by the sovereign's demand for assets, mainly in the form of safe assets. But the first-order macroeconomic implications of this shortage could be explained without the additional subtlety of isolating various risk characteristics or identifying the particular assets that were in chronic scarcity. The distinction, however, became increasingly important over time,

first covertly, then overtly in the aftermath of the subprime mortgage crisis and its sequels.

To start, the shortage of safe assets in the 2000s distorted the incentives of the financial system, especially in the United States, toward the issuance of "private label" safe assets: specifically, an explosion of the supply of AAA-rated securitized instruments manufactured by the financial industry (for example, using collateralized debt obligations based on mortgage-backed securities). Simultaneously, it made it easy for fiscally weak sovereigns such as Greece or Italy to issue debt at favorable yields. These additional assets, initially perceived as "safe" by naive investors, reduced the safe asset shortage and the downward pressure on global real interest rates. But when the subprime and European sovereign debt crises eventually erupted, the sudden loss of safe status of these pseudo-safe assets abruptly accelerated the underlying trend by simultaneously contracting the supply and increasing the demand for safe assets as most economic agents tried to de-lever. Safe interest rates declined precipitously, but soon reached their *effective lower bound*, that is, the rate at which cash becomes more attractive than financial assets and cannot be lowered further.[1]

In this analysis, the effective lower bound is a tipping point for the global economy. Any further intensification in the shortage of safe assets has destabilizing macroeconomic consequences: with safe real rates finding increasing resistance to further downward adjustment, the global economy is pushed below its potential, and the corresponding decline in global output and wealth decreases the relative demand for safe assets. This shift resorbs the safe asset shortage and restores equilibrium in the safe asset market.

This tipping point was quickly reached at the onset of the last financial crisis and contributed to the severity of the Great Recession. Today, interest rates in safe-assets-producing countries remain at or close to the effective lower bound, with very limited scope for large additional declines. The safe asset shortage remains a key source of fragility for the global economy.

In this article, we begin by describing the main facts and macroeconomic implications of safe asset shortages. Faced with such a structural conundrum, what are the likely short- to medium-term escape valves? We analyze four of them: 1) a valuation rise through the exchange rate appreciation of safe asset producer economies, and the US dollar in particular; 2) the issuance of public debt; 3) the production of private safe assets; and 4) changes in regulatory frameworks, global risk sharing, as well as re-profiling of central bank asset purchase practices to reduce the demand for safe assets. Each of these comes with its own macroeconomic and financial trade-offs, which we discuss.

[1] As is well-known, this effective lower bound is not necessarily equal to zero since storage and transportation costs may make cash unattractive even when interest rates are slightly negative. More importantly, there are many reasons besides the standard cash–bonds substitution one for why rates are difficult to reduce from very low levels: as one example, see Brunnermeier and Koby (2016) on the "reversal rate," defined as that rate below which further reductions cause more harm to the financial system than they benefit aggregate demand.

Table 1
A List of Safe Assets—Pre- and Post-Crisis

	Billions of US$		*% of world GDP*	
	2007	*2011*	*2007*	*2011*
US Federal government debt held by the public	5,136	10,692	9.2	15.8
Held by the Federal Reserve	736	1,700	1.3	2.5
Held by private investors	4,401	8,992	7.9	13.3
GSE obligations	2,910	~~2,023~~	5.2	~~3.0~~
Agency-and GSE-backed mortgage pools	4,464	~~6,283~~	8.0	~~9.3~~
Private-issue ABS	3,901	~~1,277~~	7.0	~~1.9~~
German and French government debt	2,492	3,270	4.5	4.8
Italian and Spanish government data	2,380	~~3,143~~	4.3	~~4.7~~
Safe assets	20,548	12,262	36.9	18.1

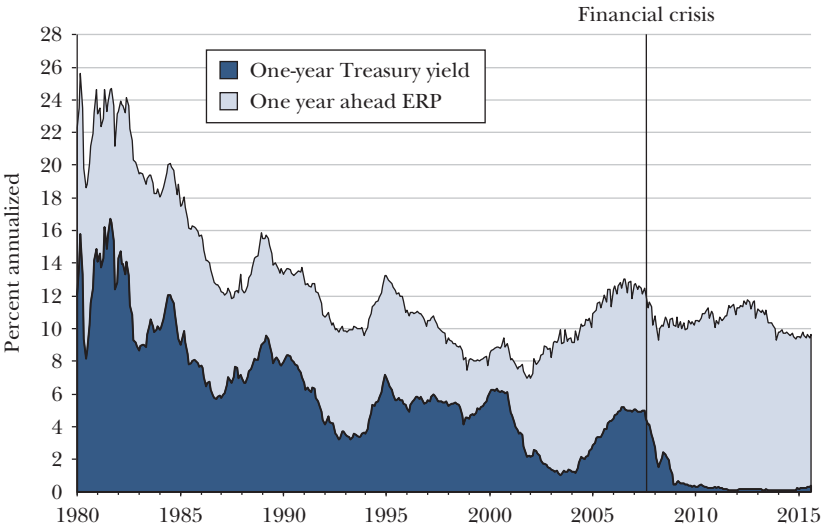
Source: Barclays Capital (2012). Data came from Federal Reserve Flow of Funds, Haver Analytics, and Barclays Capital.
Note: Numbers are struck through if they are believed to have lost their "safe haven" status after 2007. GSE means "government-sponsored enterprise." ABS means "asset-backed security."

Safe Asset Shortages and Their Macroeconomic Consequences

There have been a number of attempts in the literature to estimate the size of the pool of safe assets. All of these use somewhat crude rules to categorize assets. Table 1 presents one such measure, which includes debt from the US, German, French, Italian, and Spanish governments, together with assets held by the US "government-sponsored enterprises" such as Freddie Mac and Fannie Mae, which were heavily invested in mortgage-backed assets and were widely perceived to have the full backing of the US government. The table illustrates the collapse in the quantity of global safe assets from 2007 to 2011. Explicit US government debt rose, but mortgage-backed debt issued by the US government-sponsored enterprises was no longer perceived as safe, and neither was debt from the Italian and Spanish governments. The global quantity of safe assets plummeted as a result. Eichengreen (2016) offers an alternative and more detailed breakdown of safe assets, in which one category includes all OECD sovereign debt rated AA or above. This measure also shows a dramatic fall in safe assets during the financial crisis.

The most direct implications of a fall in the supply of safe assets can be seen in Figure 1. The two black lines in Figure 1 illustrate the paths of the short-term interest rate (dark area) and of the expected return on equity (area under the top line). The difference between the two lines is the equity risk premium (light area). Short-term rates feature a widely noted downward secular trend and a sharp drop during the Great Recession. The evolution of the expected return on equity is markedly different. It features the same downward trend as the short-term interest rate until the early 2000s, then remains more or less stable. The disconnect between a stable expected return on equity and a declining short-term interest rate is particularly salient after 2002, and even more so since the beginning of the Great Recession,

Figure 1
US Interest Rate and Expected Equity Risk Premium (ERP)



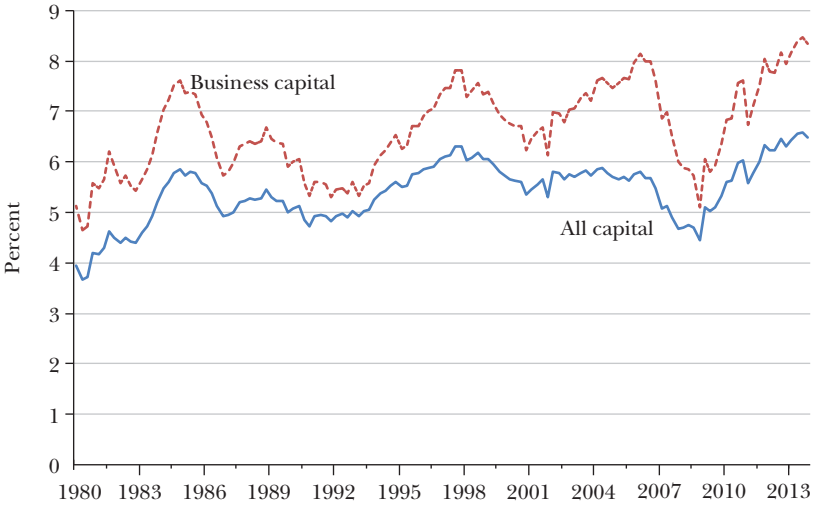
Source: One-year Treasury yield: Federal Reserve H.15; ERP: Duarte and Rosa (2015).
Note: The graph shows the one-year US Treasury yield (dark area) and the one-year expected risk premium (ERP) (grey area), calculated as the first principal component of 20 models of the one-year-ahead equity risk premium. The figure shows that the equity risk premium has increased, especially since the Global Financial Crisis.

as the latter combined a greater demand for safety and a diminution in the quantity of what were perceived as safe assets.[2] It suggests a shift towards safe assets and away from riskier ones. Figure 2 documents that over the same time period, estimates of the return to physical capital remained remarkably stable. This implies that a similar disconnect is observed between returns to capital and safe interest rates, which can also be in large part attributed to an increase in risk premia attached to physical investment (Caballero, Farhi, and Gourinchas 2017).[3]

While the underlying trend towards safe assets may have been gradual throughout the 1990s and 2000s, it was partially masked by the rapid increase in the supply of pseudo-safe assets, privately engineered by the US financial sector, as well as the increase in debt issuance by fiscally weak sovereigns such as Italy or Greece.

[2]Less consistent with the persistence of the "safety" premium is the fact that within fixed income, some credit-spreads have compressed significantly. Our conjecture is that this within-asset-class phenomenon is the result of search for yield among those intermediaries constrained by mandates and regulations rather than by their own demand for safety. It is also the kind of situation that can lead to sharp spikes in risk spreads during risk-off scenarios.
[3]Similarly, Del Negro, Giannone, Giannone, and Tambalotti (2017) find supportive evidence that the decline in safe real interest rates in the United States was driven mostly by an increase in the premium for safety and liquidity of short-term Treasury bills relative to less-liquid and less-safe assets.

Figure 2
The Average Real Return to US Capital



Source: Real after-tax returns to business capital and all capital, computed by Gomme, Ravikumar, and Rupert (2011) and adjusted for the share of intangibles in total capital from Koh, Santaeulàlia-Llopis, and Zheng (2016).
Note: The real after-tax return to capital is constructed as total after-tax capital income, net of depreciation divided by the previous period's value of capital. Business capital includes nonresidential fixed capital (structures, equipment, and intellectual property) and inventories. All capital includes business capital and residential capital.

As noted earlier, because such assets were considered safe by naive investors, this reduced the downward pressure on safe real rates.

When the financial crisis arrived in 2007–2008, the safe asset scarcity resurfaced with a vengeance. The collapse in the supply of safe assets and the increase in the demand pushed down the natural safe rate—that is, the short-term real rate required for full employment—well below zero. But nominal interest rates were already quite low and central banks around the world quickly found themselves unable to decrease nominal or real rates further. With real safe rates unable to decrease so as to clear markets, the demand for safe assets remained too elevated and the economy had to slow down and operate below its potential. This is a modern version of the paradox of thrift: faced with elevated safe real rates (relative to their equilibrium level), households prefer to save and postpone consumption; simultaneously, faced with low demand and elevated risk premia, firms prefer to postpone investment. Aggregate demand suffers and a recession ensues. In short, unable to clear markets via prices (the safe real rate), the economy clears by adjusting quantities (Caballero and Farhi 2017; Caballero, Farhi, and Gourinchas 2015, 2016).

An acute shortage of safe assets creates a situation similar to a liquidity trap, which we dub a "safety trap." Unlike the safety trap, a liquidity trap corresponds

to a situation of excess savings *across* asset classes. In both types of traps, real rates cannot fall sufficiently, causing a recession. There are, however, two important differences between safety traps and liquidity traps. First, exiting a safety trap requires an increase in the supply of, or a reduction in the demand for, safe assets, *regardless of the demand and supply of other assets*, while the more general liquidity trap calls for a reduction in saving or a general increase in stores of value. From a policy perspective, this implies that government policies that leave the supply of safe assets unchanged will be less effective (an issue we will revisit below in some detail). Second, safety traps can be very persistent or even permanent despite the presence of long-lived assets, because the risk premia attached to long-lived assets bounds the value of these assets and the associated wealth effects on aggregate demand, even with persistently low interest rates.[4]

When prices have some degree of flexibility, safety traps and the resulting recessions or periods of sluggish growth can also trigger deflationary forces, which raise real safe interest rates, further depressing output, in a familiar deflation cycle (Eggertsson, Mehrotra, Singh, and Summers 2016; Eggertsson, Mehrotra, and Robbins 2017; Caballero and Farhi 2017; Caballero, Farhi, and Gourinchas 2015).

In the discussion so far, we have considered the world economy as a single unit. The dynamics between net safe asset producers and safe asset absorbers adds substantial richness to the picture. In an open economy, the scarcity of safe assets in one country spreads to others via capital outflows, until safe rates are equalized across countries. As the global scarcity of safe assets intensifies, the global safe interest rate drops and capital flows increase to restore equilibrium in global and local safe asset markets. Once the zero lower bound for global interest rates is reached, global output becomes the adjustment variable. The world economy enters a regime of increased interdependence, since countries can no longer use monetary policy to insulate their economies from world capital flows (Caballero et al. 2015). A country with an acute scarcity of safe assets spreads its recession to other countries via capital outflows, or equivalently, current account surpluses. Surplus countries (like the eurozone) are exporting their weak domestic aggregate demand. Deficit countries, like the United States, are absorbing the weak domestic aggregate demand of the rest of the world.

The global economy can remain fragile for long periods of time, even if some countries like the United States and the United Kingdom have managed to largely erase their output gaps over time, since any intensification in safe asset scarcity in some countries could lead to the re-emergence of a global safety trap.[5]

In summary, the world economy seems to have transitioned to an environment of recurrent global safety traps: we might emerge from one, only to relapse at the

[4] In a permanent liquidity trap, in the absence of risk premia, the value of long-lived assets would become arbitrarily large as interest rates fall to zero, increasing the supply of assets and eliminating any asset shortage.

[5] According to the IMF *World Economic Outlook* (April 2017), the 2016 output gap for the UK economy was –0.17 percent while that of the US economy was –0.42 percent. The output gaps for the eurozone and Japan were estimated at –0.7 percent and –1.71 percent, respectively.

next wave of economic bad news. The next four sections explore market and policy mechanisms that may reduce, perhaps temporarily, the underlying imbalance.

Foreign Exchange Appreciation of the Currencies of Safe Asset Producers

The main market mechanism to restore equilibrium in a safety trap is an increase in the valuation of safe assets, a process that is hampered by limits on how low rates can go. However in a global economy, there is a second valuation channel: the exchange rate. An appreciation of the currency in which these assets are denominated, primarily US dollars, increases the real value of these assets for non-US holders. However, to absorb the trend increase in the net demand for safe assets, the currency of safe asset issuers needs to appreciate at a rate at least equal to the difference between the rate of growth of non-issuers and that of issuers.

The central problem of this particular "solution" is that it depresses net exports, and potentially output, for safe asset issuers. While consumers in these countries enjoy the ongoing revaluation of their income in terms of greater buying power of foreign-made goods, domestic producers experience all the burden of adjustment. In Caballero et al. (2015), we refer to this phenomenon as the *paradox of the reserve currency*.

When equilibrium full-employment interest rates in an economy are well above the effective lower bound, a reserve currency status for countries that issue safe assets is mostly an economic blessing as it allows for lower funding costs (Gourinchas and Rey 2007). But when the global economy nears the effective lower bound, safe asset issuers, faced with a wave of foreign investors seeking to invest in safe assets, will find that their currency tends to appreciate, exacerbating their own safety trap. In this setting, being the issuer of a reserve currency in which safe assets are denominated becomes a disadvantage.

This perspective also has policy implications within the set of safe-asset-producing economies. When interest rates are constrained above the equilibrium full-employment real interest rate and global output needs to decline as a result, the distribution of this global recession across countries depends on the exchange rate. By depressing the value of their currency, countries can stimulate their economy at the expense of their trading partners. This creates fertile grounds for "beggar-thy-neighbor" devaluations achieved by direct interventions in exchange rate markets, which stimulate output and improve the current account in one country at the expense of the others. The recent evolution of currency values for advanced economies illustrates this pattern. The accommodating monetary policy of the United States from 2008–2014 was associated with a substantial depreciation of the US dollar, which helped to reduce US current account deficits. In turn, the Bank of Japan (in 2013) and the European Central Bank (late 2014) launched large-scale asset purchase programs, that contributed to the depreciation of the yen and the euro against the dollar, shifting the adjustment burden back onto the US economy.

Beyond exchange rates, the general principle is that at very low interest rates, safe assets acquire a public good dimension since their production helps stimulate output in other countries, and these benefits are unlikely to be fully internalized by the economy that is issuing the safe assets. A free-rider problem arises, which manifests itself both in quantities (under-issuance of safe assets) and in prices (beggar-thy-neighbor devaluations). The US economy has clearly experienced the short end of this bargain in recent times.

The Role of Public Debt and Infrastructure Investment

One obvious solution to a shortage of safe assets is for countries that produce safe assets to issue more of them. This solution seems feasible as long as the cost of servicing public debt remains negligible, which is to say as long as interest rates stay well below the issuer's rate of growth. However, this solution is also potentially fragile and even bubble-like, because it is susceptible to rollover risk. More specifically, it requires taking the risk of becoming exposed to a coordination-failure-type run on public debt (Farhi and Maggiori 2016), or to the exploding debt dynamics that might follow a sudden decline in the demand for safe assets. While the shortage of safe assets creates space for more debt issuance than in other environments, in practice this margin is likely to be limited.

The capacity of a government to issue more safe public debt depends on two factors: the fiscal capacity of the government to borrow, and the risk that increased provision of public safe assets may crowd out provision of private-sector safe assets. In a situation with a shortage of safe assets, the relevant form of fiscal capacity is the government's perceived ability to commit to raising future taxes, even if the economic crisis were to last for a long time or worsen. On the other hand, the risk of crowding out private safe assets depends on how much these anticipated future taxes reduce the private sector's capacity to issue safe claims backed by risky dividends. Naturally, crowding out of private-sector safe assets is less likely when the securitization capacity of the economy is already impaired—since few private-sector safe assets can be constructed at such a time. In a safety trap, issuing additional public debt increases the supply of safe assets and stimulates the economy.

A substantial share of the contraction in the supply of safe assets from 2007 to 2011, shown in Table 1, resulted from a perceived violation of the fiscal capacity condition of some large eurozone economies. Since then, other economies, and the same economies with external backing, have been rebuilding this supply of safe assets.

The macroeconomic desirability of an expansion in public debt during a time of safe asset shortage is distinct from (and complementary with) the more conventional advocacy for (cheaply funded) fiscal expansion during liquidity traps and/or secular stagnation situations. The mechanism operates through a swap of risky for riskless assets in private sector portfolios. In this sense, policies that increase the gross supply of safe assets—such as "helicopter drops" of money, safe public debt issuances, and versions of central bank's quantitative easing involving swaps of "positive-beta"

private risky assets for "zero- or negative-beta" public safe assets—stimulate aggregate demand and output. Recent examples of such policies include the so-called QE1 episode in the United States from December 2008 through March 2010 and the long-term refinancing operation that started in late 2011 in the eurozone.[6]

In contrast, "Operation Twist" type policies involving swaps of "negative-beta" long-term government debt for "zero-beta" short-term public debt are ineffective or even counterproductive (Caballero and Farhi 2017). Examples of these policies would include the QE2 policy enacted by the Federal Reserve from November 2010 to June 2011 and the following QE3 policy from September 2012 to December 2013.

In a globally connected economy, an expansion in the supply of safe assets from the public debt issuance of core economies spreads across the world economy. This raises the concern that the quantity of safe assets issued by core economies and necessary to fulfill a growing global demand may be too large and eventually weaken the fiscal capacity of core economies.

While the experience of Japan over recent decades suggests that the capacity of a core economy to issue debt may be extremely large in a safe asset scarcity environment, it is a concerning situation. Again, this overall phenomenon is secular, to the extent that core economies naturally grow at a slower pace than emerging markets economies, which are heavy net users of safe assets. This is also compounded by a series of demographic factors that are increasing the demand for safe assets and reducing the effective tax-base for safe asset issuers.

This situation is a modern version of the old "Triffin dilemma," an argument made by economist Robert Triffin in various writings in the early 1960s. The original Triffin dilemma referred to the tension between the growing global demand for US dollars under the Bretton Woods system of fixed but adjustable exchange rates and a constant dollar price of gold, and the (largely) fixed amount of gold reserves held by the US government. The dilemma was that either the US monetary authorities would have to tighten monetary policy eventually, holding down the demand for US dollar assets but also causing a global recession, or the United States would find itself unable to back the stock of dollars with gold reserves, which would eventually make the system of fixed exchange rates unsustainable—as eventually happened. In the modern version of the Triffin dilemma, the demand for safe asset debt from certain countries grows with the world economy faster than the issuer's own economy (Gourinchas and Rey 2007; Farhi, Gourinchas, and Rey 2011; Obstfeld 2011; Farhi and Maggiori 2016). Expanding issuers' public debt in line with global demand runs the risk of exhausting fiscal capacity, or of a coordination failure type run on their debt. Moreover, should the environment change and the safe asset scarcity disappear, issuers could rapidly face exploding and unsustainable debt dynamics.

[6]In this context, "beta" refers to the extent to which the return on a financial asset is correlated with other returns in the market. A safe asset as defined here should have a beta of nearly zero—that is, the value of the asset should not change (much) depending on whether other assets are experiencing rising or falling returns.

Nevertheless, the issuance of safe assets has a public good dimension: production of safe assets *anywhere* expands output *everywhere* and the positive spillovers are unlikely to be fully internalized. Thus, while safe asset issuers may rightly be concerned about the risks of exhausting their fiscal capacity, the current environment is likely one of under-issuance of public debt by core economies.

In the meantime, financial engineering, such as the pooling of risks among quasi-safe sovereigns to create a larger share of safe debt from *existing* public assets can add another layer of supply of safe assets. The overall approach here involves tranches: that is, it combines a number of risky assets into a pool, then creates a series of derivative assets. The most "junior" tranche of these assets bears all of the losses, up to a certain percentage of the total. Intermediate (or mezzanine) tranches bear losses above that amount. The most senior tranches are the safest, because they only bear losses after all the lower tranches have been wiped out. Of course, junior tranches also need to offer a higher rate of return to offset their greater risk, while the most senior and safest tranches pay the lowest rate of return. This general approach is a key component of various proposals to group together sovereign bonds issued in euros. The proposal for "European Safe Bonds" (ESBies) is one prominent example (Brunnermeier et al. 2016), where liabilities for individual bonds remain with each sovereign, but the pooled assets issue a union-wide safer senior tranche. Another example is for the IMF to oversee joint tranching of emerging-markets debt as proposed in Caballero (2003), where sovereigns also keep individual liabilities but the pooled-assets issue a hard-currency safe asset tranche.[7]

From this perspective, publicly funded infrastructure investment becomes particularly attractive, as it both boosts growth in the asset-producing countries, increasing fiscal capacity, and does so with maximum issuance of safe asset per unit of installed capital.

Could other sovereign safe asset issuers come on line to add significantly to the existing, primarily US-based, supply of safe assets? As an historical precedent, the *Economist* (2015) mentions the passing of the safe asset baton from the United Kingdom to the United States in the 1930s. One could imagine an expansion in the supply of Chinese safe assets in this context, but this is probably a few decades away from becoming a significant factor. Furthermore, even if it did, the benefits of the emergence of another major issuer could be mitigated by a rise in self-fulfilling instability arising from coordination problems as investors substitute away from one issuer and into another (Nurkse 1944; Farhi and Maggiori 2016).

Private Substitutes

If the public sector is unable to expand the production of safe assets, the private sector will face powerful incentives to increase their issuance, as it did

[7]Other eurozone proposals that use tranching include the Blue Bonds/Red Bonds of Delpla and Weiszacker (2010) and the collateralized debt obligation proposal of Corsetti et al. (2016).

in the past. Private substitutes can take many forms. For example, corporations have an incentive to make themselves a safer source of return, for instance by withdrawing from investing in risky projects and instead distributing a stable dividend or buying back their own shares—both patterns that have been observed in recent years.

However, the closest private sector alternative to sovereign safe assets arises from the private sector's incentive to financially engineer substitutes. Over time, there has been a dramatic structural transformation of the composition of privately produced "safe" assets (as documented by Gorton, Lewellen, and Metrick 2012). In the early 1950s, demand deposits at banks were a safe asset. As the financial sector became more sophisticated, this category of "safe" assets expanded to include money-like debt (for example, commercial paper, money market funds, repurchase agreements) and private label AAA asset-backed securities. What is relatively new, relative to post–World War II history, is that the global economy is going through a complex structural period where the standard valuation adjustment for safe assets—via interest rate changes—have run out their course.

While it is possible in principle to create private-sector safe assets with sufficient overcollateralization (Hall 2016), this solution remains fragile since the private sector's ability to insure against a truly systemic event is limited (Holmström and Tirole 1998). In fact, much of the initial impetus behind the subprime crisis resulted from the financial sector trying to extract a (seemingly) safe asset tranche from pooled lower-quality assets (for discussion of this topic, see Caballero 2010; Stein 2012; Gorton 2016). But even the most senior and seemingly safe tranches on private assets may contain some irreducible tail-risk, making these assets unsafe when faced with truly systemic events. This creates substantial instability in the absence of an explicit public insurance overlay (Caballero and Kurlat 2009; Stein 2012). Indeed, as we argued earlier, this particular private sector attempt to create safe assets played a significant role in pushing the world economy into the Great Recession.

In essence, the financial sector in the lead-up to the financial crises was able to create *micro*-AAA assets from the securitization of lower-quality assets. But the industry remained largely unprotected against a truly systemic event, and the complexity of the instruments made them vulnerable to a panic, which duly took place. Overcollateralization does not solve such a problem: complex private safe assets are not truly robust against the potentially chaotic unraveling that follows a systemic panic, in the absence of an explicit public backstop. That is, private safe assets are not *macro*-AAA assets. As Gorton (2016) lucidly writes:

> And leading up to the recent crisis there was a shortage of long-term safe debt, so agents were increasingly using privately-produced long-term debt, AAA/Aaa asset-backed and mortgage-backed securities (ABS/MBS). The outcome of this … was the financial crisis … So, now more attention is paid to safe assets . . . This is as it should be because almost all human history can be written as the search for and the production of different forms of safe assets.

Stein (2012) expresses a similar concern about short-term liabilities created by the banking industry, which are supposedly safe but may also be subject to tail risk. His proposal to assure the safety of these assets is that the US Treasury floods the market with short-term debt. While this may improve the overall supply of safe assets temporarily, the structural problem will not be remedied, but merely postponed until the political-fiscal capacity is reached, as discussed above.

Some form of private–public partnership could help expand the private supply of safe assets. For example, one alternative would be providing fiscal backstop for the severe tail risks of safe private assets, while monitoring collective moral hazard (for example, Caballero and Kurlat 2009; Farhi and Tirole 2012). Caballero and Kurlat (2009) suggest that banks would continue with their role in the provision of safe short-term assets, but would be required to buy tail-risk macroeconomic insurance from the government. It is obvious that the design of such a program raises difficult questions. It is extremely hard, and for that reason inefficient, for the private sector alone to produce tail-risk systemic insurance. Conversely, it seems highly inefficient for the public sector to use its political debt capacity insuring nonsystemic events that can in principle be handled by the private sector. Of course, a public sector backstop also raises the question of the fiscal capacity of the government to honor that backstop, when and if needed.

Reducing the (Net) Demand for Safe Assets

If expanding the production of safe assets sufficiently is difficult, could we find areas in which safe asset demanders might be encouraged to hold fewer of them? The first area that comes to mind is the enormous pool of safe assets on central banks' balance sheets. For example, of the $18 trillion of outstanding US Treasuries, the quintessential liquid safe asset, more than 30 percent is stationed at central banks—two-thirds at foreign central banks and one-third at the Federal Reserve itself. Overall, the total assets of major central banks around the world rose from roughly $6 trillion in 2008 to $16.3 trillion by 2016. Finding alternative—if necessarily riskier—assets for central banks to hold could help to address the safe asset shortage.

Central banks hold safe assets for two main reasons: 1) to be able to intervene in foreign exchange markets if desired, which typically involves hoarding of foreign safe assets; and 2) as a result of quantitative easing policies, which involves the accumulation of domestic safe assets and occasionally riskier ones. (Although in some countries, like Japan, the policies of foreign exchange market intervention and quantitative easing can become mixed at times).

The accumulation of safe assets in the form of foreign exchange reserves, especially in emerging markets, reflects in part a precautionary motive against the occurrence of a sudden stop of capital inflows: that is, the holdings of safe assets by these central banks is for self-insurance purposes. This is an an inefficient mechanism of systemic insurance, which could be partially replaced by more powerful global risk sharing arrangements including swap lines, credit facilities backed by

international financial institutions like the IMF or the World Bank, and reserve sharing agreements (for a discussion, see Caballero 2003; Farhi, Gourinchas, and Rey 2011; Farhi and Maggiori 2016). The IMF and the Federal Reserve implemented some of these policies with foreign central banks during the peak of the financial crisis. The Federal Reserve's swap lines were credited with limiting the spreading of the US subprime crisis to the rest of the world as foreign banks that had funded themselves in dollars ran into trouble.

A central bank holding safe assets as a result of quantitative easing policies faces a very different situation. After all, quantitative easing is not an insurance policy. It is a policy adopted after an adverse economic event has already occurred, designed to compress risk-spreads. As such, it is not clear at all that it needs to involve the purchase of safe assets. In fact, if a shortage of safe assets is the main reason behind the economic downturn, and the constraints on those that demand these assets to shift their portfolios into riskier assets are severe, reducing the available supply of safe assets via central bank purchases may aggravate the problem. In that situation, it makes more sense for a central bank engaged in quantitative easing to purchase riskier assets, such as the mortgage-backed securities purchased by the Federal Reserve, or even riskier assets such as the equity shares and real estate bonds purchased by the Bank of Japan in its quantitative easing program.

To sum up, reducing safe asset hoarding by emerging and advanced economies' central banks may require different steps. For emerging markets, holding safe assets issued by a limited number of high-income countries, it requires alternative forms of global pooling of macro-risks. For high-income countries, whose central banks hold a substantial share of the world's safe assets, it requires consideration for the policy spillovers of the different quantitative easing options available. Put differently, in the current environment, developed markets' central banks should not be hoarding assets that have a large safe asset component beyond those required for the conduct of conventional monetary policy.

A final area in which the demand for safe assets might be reduced involves some rethinking of the regulatory framework. Flow of funds data indicate that one key source of the global demand for safe dollar assets originates within the global financial sector (Gorton, Lewellen, and Metrick 2012; Gourinchas and Jeanne 2012). Well-intentioned but perhaps shortsighted new regulatory requirements implemented in the aftermath of the financial crisis have significantly increased the mandated safe asset holdings of financial institutions, especially banks and insurance companies, under the Basel III criteria currently being phased-in. Finding ways to safeguard the stability of the financial sector without generating high demand for safe-assets would also alleviate the scarcity.

Taking Stock

In the short- and medium-run, the world economy is likely to remain unpleasantly close to a structural safety trap, unless some powerful steps are taken. As

Gorton, Lewellen, and Metrick (2012) showed, the share of safe assets relative to total US assets has been remarkably stable over the long-run of recent decades, suggesting that the long-run trend toward increased scarcity of safe assets has been mostly due to demand factors, such as central banks' international reserve accumulations, regulatory changes, and demographic factors.

The ongoing pressures driving the imbalance in safe asset markets has in recent decades helped to drive the steady decline in interest rates on safe assets. However, interest rates on these assets cannot fall much further. When the equilibrium full-employment interest rate needs to be negative, but cannot adjust sufficiently downward, then (other things equal) the equilibrating mechanism is an endogenous decline in safe asset demand through a reduction in aggregate income and wealth. That is, equilibrium is achieved through recession.

Another top-down way of thinking about the general macroeconomic malaise caused by the shortage of safe assets is to consider the physical investment that is required to match the saving needs of society. If the saving side of society has a disproportionate desire for safe assets, then it effectively wants to fund only a small share of the overall risky investments required for economic growth. A central role played by the financial sector is to intermediate risk between the savers who want safe assets and the borrowers who are taking on a greater degree of risk. One result of such intermediation is that the interest rates associated with the relatively small tranches of safe assets are compressed against the zero lower bound, while other risk spreads remain elevated. If the financial sector cannot fully manage this transmutation, then it will be hard to sustain the levels of physical investment needed to generate growth in core economies—and it will be hard for these core countries to carry out an ongoing expansion of the quantity of safe assets. From this perspective, publicly funded infrastructure investment becomes particularly attractive, as it both boosts potential growth in the asset producer countries and does so with maximum issuance of safe assets per unit of installed capital.

In the short- and medium-term, the quantity of safe assets may increase via stronger exchange rates in the safe asset issuers, and via public debt issuance in those countries. Over time, a lasting solution to the shortage of safe assets will require a combination of finding alternative sources of safe asset supply and a reduction in demand. Some years down the road, current emerging markets, especially China, may eventually provide global safe assets in substantial quantities, but for now, this avenue holds little promise. Reconsidering how and why central banks hold safe assets as reserves and as part of quantitative easing, and also rethinking the rules that require private financial firms to hold safe assets, are potentially ways to increase the quantity of safe assets available in the market. In the meantime, as the global economy struggles to find ways to reduce the shortage of safe assets, we are likely to continue to see the multiple symptoms of this economic illness: very low interest rates on safe assets, bubbly expansions of seemingly safe assets, recessions, episodic sharp appreciations of core currencies, and so on.

■ *We thank the journal's editors Enrico Moretti, Mark Gertler, Gordon Hanson, and Timothy Taylor as well as Marion Fourcade for insightful comments.*

References

Barclays Capital. 2012. *Equity Gilt Study 2012.* Available at: http://topforeignstocks.com/2012/10/09/download-barclays-equity-gilt-study-2012/.

Bernanke, Ben S. 2005. "The Global Saving Glut and the U.S. Current Account Deficit." Lecture presented at the Sandridge Lecture, Virginia Association of Economists, Richmond, Virginia, March 10.

Brunnermeier, Markus K., and Yann Koby. 2016. "The Reversal Interest Rate: The Effective Lower Bound of Monetary Policy." Unpublished paper, Princeton University.

Brunnermeier, Markus K., Sam Langfield, Marco Pagano, Ricardo Reis, Stijn Van Nieuwerburgh, and Dimitri Vayanos. 2016. "ESBies: Safety in the Tranches." Paper presented at the 65th Economic Policy Panel, Florence, Italy, March 14–15.

Caballero, Ricardo J. 2003. "The Future of the IMF." *American Economic Review* 93(2): 31–38.

Caballero, Richardo J. 2006. "On the Macroeconomics of Asset Shortages." In *The Role of Money—Money and Monetary Policy in the Twenty-First Century,* edited by Andreas Bayer and Lucrezia Reichlin, 272–83. Frankfurt: European Central Bank.

Caballero, Ricardo J. 2010. "The 'Other' Imbalance and the Financial Crisis." NBER Working Paper 15636.

Caballero, Ricardo J., and Emmanuel Farhi. 2017. "The Safety Trap." *Review of Economic Studies.* Accepted paper: http://www.restud.com/wp-content/uploads/2017/02/MS20803manuscript.pdf.

Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2008. "An Equilibrium Model of 'Global Imbalances' and Low Interest Rates." *American Economic Review* 98(1): 358–93.

Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2015. "Global Imbalances and Currency Wars at the ZLB." NBER Working Paper 21670.

Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2016. "Safe Asset Scarcity and Aggregate Demand." *American Economic Review* 106(5): 513–18.

Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2017. "Rents, Technical Change, and Risk Premia Accounting for Secular Trends in Interest Rates, Returns on Capital, Earning Yields, and Factor Shares." *American Economic Review* 107(5): 614–20.

Caballero, Ricardo J., and Pablo Kurlat. 2009. "The "Surprising" Origin and Nature of Financial Crises: A Macroeconomic Policy Proposal." Paper presented at the Financial Stability and Macroeconomic Policy Federal Reserve Bank of Kansas City Symposium, Jackson Hole, WY, August 20–22.

Caballero, Ricardo J., and Alp Simsek. 2013. "Fire Sales in a Model of Complexity." *Journal of Finance* 68(6): 2549–87.

Corsetti, Giancarlo, Lars P. Feld, Ralph Koijen, Lucrezia Reichlin, Ricardo Reis, Hélène Rey, and Beatrice Weder di Mauro. 2016. "Reinforcing the Eurozone and Protecting an Open Society." London: CEPR Press.

Dang, Tri Vi, Gary Gorton, and Bengt Holmström. 2015. "Ignorance, Debt, and Financial Crises." http://www.columbia.edu/~td2332/Paper_Ignorance.pdf.

Del Negro, Marco, Domenico Giannone, Marc P. Giannoni, and Andrea Tambalotti. 2017. "Safety, Liquidity, and the Natural Rate of Interest." *Brookings Papers on Economic Activity,* no. 1.

Delpla, Jacques, and Jacob von Weizsacker. 2010. "The Blue Bond Proposal." *Bruegel Policy Brief* 3: 1–8.

Duarte, Fernando, and Carlo Rosa. 2015. "The Equity Risk Premium: A Review of Models." Staff Report 714, Federal Reserve Bank of New York.

Economist, The. 2015. "Dominant and Dangerous." October 3. http://www.economist.com/news/leaders/21669875-americas-economic-supremacy-fades-primacy-dollar-looks-unsustainable-dominant-and.

Eichengreen, Barry. 2016. "Global Monetary Order." In *The Future of the International Monetary and Financial Architecture* (Conference proceedings, June 27–29, Sintra, Portugal), edited by Vitor Constancio and Philipp Hartmann, 21–63. European Central Bank.

Eggertsson, Gauti B., Neil R. Mehrotra, and Jacob A. Robbins. 2017. "A Model of Secular Stagnation: Theory and Quantitative Evaluation." NBER Working Paper 23093.

Eggertsson, Gauti B., Neil R. Mehrotra, Sanjay R. Singh, and Lawrence H. Summers. 2016. "A Contagious Malady? Open Economy Dimensions of Secular Stagnation." *IMF Economic Review* 64(4): 581–634.

Farhi, Emmanuel, Pierre-Olivier Gourinchas, and Hélène Rey. 2011. *Reforming the International Monetary System*. London: CEPR.

Farhi, Emmanuel, and Matteo Maggiori. 2016. "A Model of the International Monetary System." NBER Working Paper 22295.

Farhi, Emmanuel, and Jean Tirole. 2012. "Collective Moral Hazard, Maturity Mismatch, and Systemic Bailouts." *American Economic Review* 102(1): 60–93.

Gomme, Paul, B. Ravikumar, and Peter Rupert. 2011. "The Return to Capital and the Business Cycle." *Review of Economic Dynamics* 14(2): 262–78.

Gorton, Gary B. 2016. "The History and Economics of Safe Assets." NBER Working Paper 22210.

Gorton, Gary B., Stefan Lewellen, and Andrew Metrick. 2012. "The Safe-Asset Share." NBER Working Paper 17777.

Gourinchas, Pierre-Olivier, and Olivier Jeanne. 2012. "Global Safe Assets." BIS Working Paper 399.

Gourinchas, Pierre-Olivier, and Hélène Rey. 2007. "From World Banker to World Venture Capitalist: U.S. External Adjustment and the Exorbitant Privilege." In *G7 Current Account Imbalances: Sustainability and Adjustment*, edited by Richard H. Clarida, 11–66. Chicago: University of Chicago Press.

Hall, Robert E. 2016. "The Role of the Growth of Risk-Averse Wealth in the Decline of the Safe Real Interest Rate." http://web.stanford.edu/~rehall/DSFIR11032016.

He, Zhiguo, Arvind Krishnamurthy, and Konstantin Milbradt. 2016. "What Makes US Government Bonds Safe Assets?" NBER Working Papers 22017.

Holmström, Bengt, and Jean Tirole. 1998. "Private and Public Supply of Liquidity." *Journal of Political Economy* 106(1): 1–40.

Koh, Dongya, Raül Santaeulàlia-Llopis, and Yu Zheng. 2016. "Labor Share Decline and Intellectual Property Products Capital." Working Papers 927, Barcelona Graduate School of Economics.

Nurkse, Ragnar. 1944. *International Currency Experience: Lessons of the Interwar Period*. Geneva: League of Nations.

Obstfeld, Maurice. 2011. "International Liquidity: The Fiscal Dimension." NBER Working Paper 17379.

Stein, Jeremy C. 2012. "Monetary Policy as Financial Stability Regulation." *Quarterly Journal of Economics* 127(1): 57–95.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 47–66

Dealing with Monetary Paralysis at the Zero Bound

Kenneth Rogoff

Despite an outward appearance of stability, the core of the global monetary system today is immersed in a level of intellectual turmoil not seen since the breakup of the Bretton Woods system in the early 1970s. Back then, it was the system of fixed exchange rates that constrained central banks (except for the United States at the center). More recently, the key constraint for central banks is the zero lower bound on nominal interest rates. The zero bound has its roots in a diverse range of frictions but is due above all to the fear of central banks that if they push the short-term policy interest rates, which they set, too deeply negative, there will be a massive flight into paper currency. Cash, of course, pays no interest, positive or negative.

This paper asks whether, in a world where paper currency is becoming increasingly vestigial outside small transactions (at least in the legal, tax-compliant economy), there might exist relatively simple ways to finesse the zero bound without affecting how most ordinary people live. Surprisingly, this topic has been relatively obscure during the past decade compared to the massive number of articles, well-represented in top journals, that take the zero bound as given and look for out-of-the-box solutions for dealing with it. In an inversion of the old joke, it is a bit as if the economics literature has insisted on positing "assume we *don't* have a can opener," without considering the possibility that we might be able to devise one.

■ *Kenneth Rogoff is Professor of Economics and Thomas D. Cabot Professor of Public Policy, Harvard University, Cambridge, Massachusetts. His email address is krogoff@harvard.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.47

doi=10.1257/jep.31.3.47

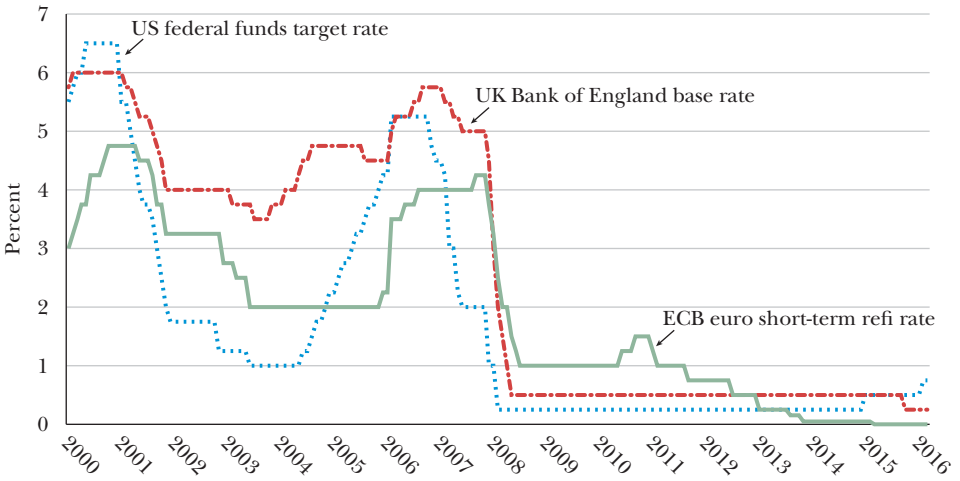
The path to effective negative interest rate policy is hardly something that can be implemented overnight. As I will argue, however, it makes sense not to wait until the next financial crisis to develop plans and, in any event, it is time for economists to stop pretending that implementing effective negative rates is as difficult today as it seemed in Keynes' time. The growth of electronic payment systems and the increasing marginalization of cash in legal transactions creates a much smoother path to negative rate policy today than even two decades ago. Fundamentally, there is no practical obstacle to paying negative (or positive) interest rates on electronic currency and, as we shall see, effective negative rate policy does not require eliminating paper currency.

There are a variety of mechanisms to avoid a run into paper currency if the central bank needs to steer short-term policy interest rates deeply negative (say, to combat a major systemic financial crisis). One involves getting rid of large-denomination notes, which Henry (1976) and myself (Rogoff 1998, 2015, 2017) argue would be a good idea for fighting tax evasion and crime regardless. The other approach does not touch cash at all, but instead creates a crawling pegged exchange rate between paper currency and bank reserves. It might sound head-spinning for a single country to have two currencies, but as we shall see, it is not as complicated as it sounds.

Of course, part of economists' fascination with the zero lower bound is precisely that it forces a rethinking of conventional dogma. Just as the laws of physics imply strange and surprising consequences as an object approaches a black hole, the laws of economics can yield some strange and surprising results as an economy gets too near the zero lower bound on interest rates. Fiscally irresponsible budget policy can become responsible, and structural reforms to make economies more efficient can become counterproductive. The foundations of the international monetary system can be threatened by a shortage of safe assets, which economists once thought impossible under flexible exchange rates. To reduce the possibility that economies will get stuck on the zero bound in future crises, a significant number of leading macroeconomists have argued central bankers should abandon all pretense of long-term price stability and raise their inflation targets to 4 percent.

Although the modeling and empirical issues are indeed very interesting for researchers, this is hardly an idle academic discussion. With today's ultra-low policy interest rates—inching up in the United States and still slightly negative in the eurozone and Japan—it is sobering to ask what major central banks will do should another major prolonged global recession come anytime soon. During nine recessions since the mid-1950s, the US Federal Reserve has cut its policy interest rate by an average of 5.5 percentage points (Yellen 2016). There is hardly room for that now, or into the foreseeable future. Yes, during the financial crisis, central banks developed a number of unconventional monetary policy tools such as "quantitative easing," but many economists are rightly concerned that unconventional monetary policy tools are poor substitutes for conventional interest rate policy and might well have more side-effects. Hence it becomes a research imperative to consider alternatives.

Figure 1
Policy Interest Rates for the Federal Reserve, European Central Bank, and the Bank of England



Source: Federal Reserve Board, European Central Bank, and Bank of England, February 2017.

The Sharp Fall in Nominal Central Bank Policy Interest Rates

The level of policy angst and research interest on how to navigate the zero bound reflects the very low interest rate environment in advanced countries that, outside of Japan, has not been seen since the Great Depression of the 1930s. It also reflects a view that even after post–financial crisis reflation, the general level of nominal interest rates is likely to remain suppressed for a long time to come. Intuitively, the lower the starting point for policy interest rates when a recession hits, the greater the odds of bumping into the zero lower bound.

Figure 1 plots policy interest rates for the United States, the European Central Bank, and the Bank of England since 2000. Notice that the Federal Reserve cut the federal funds interest rate target by roughly 5 percentage points after the bursting of the tech bubble in 2000, eventually falling to a level of 1 percent in 2003. The Fed subsequently tightened monetary policy, but then cut rates again by 5 percentage points to an official target range of 0 to 0.25 percent as the recession and global financial crisis unfolded in 2007–2008.

The European Central Bank does not have the Fed's long track record, having faced only two recessions since its founding in 1999. However, the ECB did cut its policy rate, the short-term euro refinancing rate, by 2.5 percentage points in the early 2000s recession and later by over 4 percentage points during the global financial crisis. As of March 2017, banks depositing funds at the ECB received –0.4 percent. The Bank of England base rate, also shown in Figure 1, follows a similar

pattern. Japan suffered its financial crisis starting in 1992 (Reinhart and Rogoff 2009), and the Bank of Japan's policy rate has been hovering around zero for roughly two decades and is now slightly negative.

The recent collapse in monetary policy interest rates to near zero is quite remarkable. When John Taylor (1993) first estimated his famous "Taylor rule" for monetary policy in 1993, he suggested that a normal central bank policy interest rate ought to be around 4 percent, which represented a combination of 2 percent target inflation rate and 2 percent "neutral" short-term real interest rate. For many years, major inflation-targeting floating-exchange-rate central banks used versions of the Taylor rule to benchmark their policies. But especially in the aftermath of the financial crisis, setting policy interest rates has not been nearly so straightforward.

Today's near-zero nominal short-term interest rates partly reflect the fact that central banks have been undershooting their inflation targets, thereby muting inflation expectations. But most of the action has come in the collapse of the equilibrium short-term real (inflation-adjusted) interest rate, which is now closer to –1 percent on average across the advanced countries than to Taylor's (1993) +2 percent (Holston, Laubach, and Williams 2016). For example, the interest rate on a 10-year inflation-indexed Treasury security fell from 2.7 percent before the financial crisis to almost –0.9 percent at the end of 2012; it rose subsequently, but by early March 2017 was still only 0.5 percent.

Several potential causes underlying this remarkable fall in real interest rates have been suggested, and there are a wide range of views on the quantitative significance of each. The variety of explanations include: increases in global savings due to the demographic cycle (Carvalho, Ferrero, and Nechio 2016); emerging-market demand for safe advanced-country assets (for example, Bernanke 2005); lower trend productivity growth (Gordon 2016); the falling cost of investment goods (Karabarbounis and Neiman 2014); and secular stagnation in world aggregate demand, perhaps exacerbated by rising inequality of incomes (Summers 2013).

There is also a strong case to be made that a good part of the recent drop in real interest rates is a legacy of the 2008 financial crisis, and of an ongoing debt supercycle that was originally centered in the United States and then in the eurozone, and now perhaps has reached China (Rogoff 2016). On top of lingering debt overhang in some regions, the financial crisis has led investors to place a greater weight on tail risks, which can in turn lead to a sharp drop in safe real interest rates even with normal risk aversion parameters (for discussion, see Reinhart, Reinhart, and Rogoff 2015, who build on Barro 2006). Indeed, options prices reveal that even though market volatility has greatly abated since the peak of the crisis, concern over tail risk remains very high. Kozlowski, Veldkamp, and Venkateswaran (2017) argue that tail risk can explain a wide range of post-crisis phenomena, not least including low investment and a large (roughly 12 percent for the United States) drop in potential output.

Another factor is that heightened post-crisis financial regulation and weak bank balance sheets have made it more difficult for small and medium-size businesses to gain access to credit markets, even controlling for slower trend growth.

Reinhart and Sbrancia (2015) argue that post-crisis, rich-country regulatory policies, which emphasize liquidity and safety cushions, have tilted the playing field in favor of sovereign borrowers. One only need look at the eurozone, where national debts have been siloed into corresponding national banks, to see an example of their idea. Geanakoplos (2014) points out that even though posted interest rates for small and medium-size borrowers can appear to be quite low, there is considerable credit-rationing for this group. Although headwinds are fading, particularly in the US economy, many potential borrowers around the world face considerably stricter collateral constraints than they did before the crisis.[1]

The implication of very low expected inflation and real interest rates is that "neutral" central bank policy interest rates are likely to remain low for many years to come. Laubach and Williams (2015) estimate a neutral nominal federal funds rate of 2 percent, well below Taylor's (1993) estimate of 4 percent. Holston, Laubach, and Williams (2016) extend the approach to look at Canada, the eurozone, and the United Kingdom and find similar declines in the neutral policy rate. Even allowing for some reversion to the mean in global real interest rates, Federal Reserve chair Janet Yellen (2016) has suggested that a neutral Fed Funds rate (a rate consistent with full employment and the Fed's 2 percent inflation target) will likely land around 3 percent.

Interest rates could surprise on the upside for any number of reasons, not least because of higher macroeconomic volatility due to a rise in populism. The central scenario, however, at least per current global bond markets, is that the general level of global interest rates is likely to remain low for some time to come, implying significant risk that central banks may have to wrestle again with a severe zero-bound episode sometime in the next couple decades, if not even the next few years.

Can Alternative Monetary Tools Obviate the Zero-Bound Constraint?

Central banks, naturally, want to reassure everyone that there is no reason to be overly concerned, and that they have already developed fully adequate alternatives to normal interest rate policy, should the need arise. These alternative tools include "forward guidance" over the path of future interest rates (with the idea of lowering today's real interest rate by raising the expectation of future inflation) and "quantitative easing" policies involving large-scale purchases of public and private bonds. Drawing on results from simulations of the Fed's empirical macroeconomic model, Yellen (2016) and Reifschneider (2016) argue that these alternative tools, already battle-tested during the financial crisis, can be fully as effective in stabilizing output and unemployment in a deep recession as being able to use the kind of unfettered negative interest rate policy discussed later in this paper. Wu and Xia (2016) take a

[1] Gourinchas and Rey (2016) show that today's low consumption/wealth ratio in the advanced world is likely a predictor of a sustained period of low global real interest rates, but not necessarily a predictor of lower trend growth.

very different path to the same conclusion, by constructing a "shadow interest rate" that attempts to take account of the overall effect of diverse Fed instruments on the economy.

Unfortunately, it is extremely difficult to produce convincing evidence, in part because experience with alternative monetary policy tools has been so limited, and results are very sensitive to modeling assumptions (Woodford 2012; Rogoff 2017). It is probably fair to say that the consensus among researchers is that the use of alternative policy instruments has probably been worth the risk but that their effectiveness has been limited. Nevertheless, before turning to an array of more radical proposals that economists have advanced, we discuss further the instruments that have already been used.

Quantitative Easing and Forward Guidance

Virtually every advanced-country central bank has engaged in some form of large-scale asset purchases, or quantitative easing, which involves issuing central bank reserves (essentially very short-term government debt) to purchase both public and private assets. The Federal Reserve engaged in quantitative easing to the tune of about 25 percent of GDP, but the Bank of Japan and the European Central Bank have done even more. The Bank of Japan's program over the past four years has been particularly aggressive, with the Bank of Japan well on track to buying public debt equal to 100 percent of GDP. Despite all its efforts, the inflation rate in Japan remains well below the Bank of Japan's 2 percent target, and long-term projections are that it may well fall even lower.

The limitations of alternative monetary instruments are underscored by the fact that the Bank of Japan has essentially tried even "helicopter money," an approach suggested by Bernanke (2002) based on Milton Friedman's famous thought experiment of having the central bank simply print money and hand it out. In fiscal year 2015, the Bank of Japan purchased far more government debt (80 trillion yen) than the government issued (30 trillion yen), and it did much the same in fiscal 2016 even after the Abe government's July 2016 announcement of a massive (28 trillion yen) new debt-financed fiscal stimulus.[2] The effects were positive but not large.

The real problem is that central banks don't have authority to make fiscal transfers (Cecchetti and Schoenholtz 2016), a nuance that seems to have largely escaped the global commentariat. Moreover, if central banks were ever to acquire the capacity to engage in helicopter money on their own, they would risk quickly losing any semblance of independence as politicians raced to use helicopter money to make opportunistic transfers.

[2] For details of the Japanese policy, see the Bank of Japan announcement of July 29, 2016, at http://www.boj.or.jp/en/announcements/release_2016/k160729a.pdf. See also the Japan Ministry of Finance plans for issuing debt at http://www.mof.go.jp/english/jgbs/debt_management/plan/e20151218issuanceplan.pdf and http://www.mof.go.jp/english/jgbs/debt_management/plan/e20160824issuanceplan.pdf.

The European Central Bank has walked a tightrope with its quantitative easing policies, because of course there is no government debt instrument for the eurozone as a whole, only national bonds. ECB quantitative easing therefore amounts to buying pro-rata shares of the debt of member states. Given that investors vastly prefer to hold German debt than, say, Portuguese or Italian debt, ECB quantitative easing policy involves actuarial transfers across governments, even if these transfers are not realized.

Another form of quantitative easing is the acquisition of private-sector stocks and bonds, which might be called "fiscal quantitative easing," because it can be decomposed into normal quantitative easing (issuance of central bank reserves to buy longer-term government debt), combined with issuance of government debt to buy private debt assets (normally viewed as directed credit). In effect, fiscal quantitative easing uses taxpayer guarantees to subsidize private companies. In theory, this tool can be very effective at the zero bound, far more effective than central bank purchases of government debt. Caballero, Farhi, and Gourinchas (2016), for example, show how global central bank purchases of risky private debt can help boost growth and prices, whereas purchases of government debt would have little effect.

The downside to directed credit is that it exposes the central bank to political pressures—for example, to buy bonds of favored sectors, companies, or financial institutions. Such measures harken to the days before financial liberalization when many European central banks, for example in France and Italy, were de facto central planners. The fear is that in today's much larger capital markets and advanced economies, fiscal quantitative easing could prove a slippery slope.[3]

Another idea is to drive down real interest rates (when nominal rates are stuck at zero) by talking up future inflation through "forward guidance." One way to do this is for the central bank to commit not to raise interest rates too quickly, even after the economy returns to full employment.[4] In principle, forward guidance to raise inflation expectations can be used to stimulate consumption and investment just as effectively as nominal interest rate cuts, since both work by lowering the real interest rate. Unfortunately, it is hard to make forward guidance credible, given

[3] Greenwood, Hanson, and Stein (2016) argue that even though the US Treasury effectively owns the Federal Reserve, the central bank can still play a helpful role in promoting financial stability by issuing very short-term bank reserves (or Federal Reserve debt) to buy up longer-term Treasury bills. They argue that the Treasury is not willing to issue at quite the same short horizons as the Fed, even though regulation gives many banks and financial market firms a strong appetite for super-short-maturity debt. However, shortening the maturity structure of debt exposes the taxpayers to greater risks to, say, a rapid and unexpected rise in global real interest rates, a risk that can hardly be ruled out given that so little is known with a high level of confidence about why interest rates fell so quickly.

[4] In Canzoneri, Henderson, and Rogoff (1983), we provide an early model of forward guidance, showing that if the central bank cannot respond by using the current interest rate (in our case due to implementation lags, rather than the zero-bound constraint), it is still possible for monetary policy to be just as effective by committing to manipulate future inflation in a way that gives the same real interest response as if the current nominal rate could be moved, as in Woodford (2012).

1) the turnover in central bank governing boards, and 2) the central bank has an incentive not to keep its promise if the economy does indeed recover.

Overall, alternative monetary policy instruments such as forward guidance and quantitative easing offer some theoretical promise for addressing the zero bound. But these policies have now been deployed for some years—in the case of Japan, for more than two decades—and at least so far, they have not convincingly shown an ability to decisively overcome the problems posed by the zero bound.

Higher Inflation Targets

A more radical idea is to raise central bank inflation targets from 2 percent to 4 percent. The idea is that if the inflation rate is, on average, 2 percent higher, the general level of nominal interest rates should be (on average) 2 percent higher as well; after all, theory teaches that monetary policy is neutral in the long run and cannot affect long-term equilibrium real interest rates. Thus, in principle, the central bank might be expected to have an extra 2 percent of nominal rate cuts to play with in a deep recession.

The pioneering papers on 4 percent inflation targets include the early quantitative analysis of Fuhrer and Madigan (1997) and the theoretical analysis of Krugman (1998). The idea really took off, though, with Blanchard, Dell'Ariccia, and Mauro (2010), written when Olivier Blanchard was Chief Economist at the International Monetary Fund. Blanchard and his co-authors argued that after the near brush with the zero bound at the beginning of the 2000s, followed by collapse to the zero bound in the global financial crisis, central banks needed to consider allowing higher trend inflation.

Raising the inflation target to 4 percent is a plausible approach, but it is not without drawbacks. First and foremost, central banks have invested over two decades in convincing the public that they are deeply committed to a 2 percent target, and that 2 percent inflation should be considered the moral equivalent of price stability. Any transition to a higher inflation target is likely to be quite disruptive, and it may never be possible to make the new higher target as credible as the old one. After all, if central banks changed their inflation target once, what is to stop them from changing their minds again?

A deeper problem, which is not simply transitional, is that there is arguably a fundamental difference between 2 percent and 4 percent inflation psychologically. At 2 percent inflation, most citizens feel little need to think much about inflation, especially as official indices likely overstate inflation due to the difficulty of incorporating new goods for which prices did not previously exist. Higher levels of inflation, if sustained for a long period, would likely lead to more indexing and more frequent price adjustment,[5] which in turn would undermine the potency of monetary policy. Simply put, central banks might find themselves needing much of the interest room

[5] Nakamura, Steinsson, Sun, and Villar (2016) find that during the high-inflation 1970s, price-setting frequency was fairly stable; nevertheless, if the higher inflation is predictable and in place for a very long period, one would strongly expect an adjustment to more frequent price setting.

accorded by a higher inflation target simply to achieve the same degree of stabilization. This would be particularly problematic in a very deep recession such as in the aftermath of a financial crisis, where inflation might collapse for a sustained period requiring very deep interest rate cuts to bring it back, implying that the bite of the zero lower bound might still be quite severe.

To the extent that the frequency of wage and price adjustment did not change, even after a long adaptation period, then higher target inflation implies greater distortion across relative prices in a world where wage and price adjustment is staggered. This effect can be quite empirically significant, as Ascari and Sbordone (2014) document in a broad-ranging study. A very important detail is that the economy must bear the relative price distortions resulting from higher inflation all the time, not just during recessions.

All in all, despite its drawbacks, the idea of raising target inflation rates is an important one, and would be well worth considering if the significantly more elegant approach of (effective) negative nominal rate policy were not available (albeit after a longer preparation period). In the meantime, concerns that problems with the zero bound might make monetary policy relatively impotent in future deep recessions has set academic researchers looking at a wide range of backup tools. The ideas are all interesting, although each comes with its own set of problems.

Implications of the Zero Bound for Broader Macroeconomic Policy Debates

It has been known since Keynes that the zero bound can increase the case for fiscal stimulus beyond what would be warranted if monetary policy were not paralyzed. However, calibrating the intensity and duration of the "excess" stimulus is far from straightforward. In his early and prescient paper on the zero bound, Lebow (1993) makes the case for leaning more on fiscal stimulus than would otherwise be warranted. A temporary fiscal stimulus that is calibrated to come off when the economy lifts off the zero bound is significantly more effective than one that lasts indefinitely, because the drag from expected future taxes is less (Christiano, Eichenbaum, and Rebelo 2011). DeLong and Summers (2012; and also Eggertsson, Mehrotra, Singh, and Summers 2016) argue that fiscal deficits can lead to *lower* debt-to-GDP ratios in a depressed economy at the zero bound, because of their effect on nominal GDP growth.

Perhaps more surprising to many economists is that a number of policies normally thought of as structural can—in a situation with the zero bound—have profound aggregate demand effects through their impact on the real interest rate, at least in principle. In a very creative and influential series of papers with various co-authors, Gauti Eggertsson has argued that when monetary policy is temporarily paralyzed by the zero bound, one has to look carefully at the price effects of any structural adjustment or macroeconomic policy. Suppose, for example, a competitiveness-enhancing reform to goods or labor markets leads over time to lower

prices through increased efficiency. Normally, any deflationary impact on aggregate demand would be a second-order issue that the monetary authorities could easily counteract by lowering interest rates. But at the zero bound this is not possible and the adverse aggregate demand effects of higher real interest rates (because of lower inflation) can have a first-order impact (for example, see Eggertsson 2010; Eggertsson, Ferrero, and Raffo 2014; Eggertsson, Mehrotra, Singh, and Summers 2016). Similarly, increased price flexibility, normally thought to make an economy more efficient, can also be problematic at the zero bound (Werning 2011).

Some of these ideas, which have been quite influential in the policy debate, are reminiscent of policies adopted during the Great Depression to fight deflation. These included suspending antitrust policies in a way that increased monopoly concentration and raised prices in a movement (at least for a time) away from the zero bound, but also arguably had a major adverse impact on the long-term path of output (Ohanian 2001).

Not every structural reform is deflationary. Nevertheless, the new literature has produced important counterexamples to the conventional wisdom that countries should always take advantage of a financial crisis to engage in politically difficult structural reforms and that those who do will typically enjoy the strongest and most durable recoveries.

One can stretch this logic to suggest that almost anything that raises expected inflation is worth considering, thanks to the positive aggregate demand effects of a lower real interest rate. Eichengreen (1986, 2016), for example, argues that when an economy is at the zero bound, it is possible that trade protectionism might prove beneficial in the short run—though not if there is retaliation by a country's trading partners. Bodenstein, Guerrieri, and Gust (2013) show that in an economy stuck at the zero lower bound, oil price increases may be much less problematic than normally presumed for oil importers, once again because the price inflation effect helps reduce real interest rates.

The zero bound also can greatly complicate international transmission of monetary policy. When one country is mired in the zero bound, it can suck in other countries as well (Caballero, Farhi, and Gourinchas 2016; Eggertsson, Mehrotra, Singh, and Summers 2016). For example, the eurozone effectively transmits lower interest rates to the rest of the world through its large current account surpluses, which sap global demand in other areas, thereby tending to drive down rates. Similarly, Japan's chronic trade surpluses have put downward pressure on US and European interest rates.

Although negative transmission of demand shocks can take place in normal times, it becomes more severe if affected countries themselves hit the zero bound, thereby losing their own capacity for countercyclical monetary policy. Farhi and Maggiori (2016) argue that the problem of the zero bound has greatly compounded a modern-day parallel to the "Triffin dilemma" that plagued the postwar Bretton Woods fixed exchange rate system.

Robert Triffin was a Belgian-American economist who at various times worked at the Federal Reserve, the IMF, the OECD, and Yale University. Triffin pointed out

that a system in which the US dollar was pegged to gold, and other countries pegged to the dollar, contained a fundamental inconsistency. As countries in the rest of the world grew, they required an increasing supply of US dollars to maintain their exchange rate pegs and conduct transactions. This implied that the United States had to issue ever-growing debts (in Triffin's analysis, through currency account deficits). But with the rest of the world growing faster than the US economy and faster than its gold supply, eventually the dollar's gold backing would lose credibility. Eventually, of course, the fixed exchange rate system did collapse and the world transitioned to floating rates, which seemed to address the problem.

The modern parallel is that with emerging markets growing faster than advanced economies, there has been a strong and rising demand for "safe" advanced-economy debt. For much of the 2000s, market clearing for advanced-country bonds has required an ever-falling interest rate. Once the zero bound becomes binding, though, a lower interest rate can no longer clear the market. Advanced economies can still meet the demand by allowing their debts to outstrip their income growth, but then the "safe assets" they issue may eventually become risky. Regardless, the fact there is so little agreement about how major central banks should deal with the zero bound creates greater uncertainty about the future and creates volatility for emerging markets, as Rajan (2016) has emphasized.

The vigorous and stimulating debate over alternative mechanisms for dealing with the zero bound is certainly fascinating. However, so many of the policy proposals are clearly second- and third-best alternatives to normal monetary policy, which begs the question of whether the zero bound is really the barrier now that most economists and policymakers still believe it to be.

Paths to Effective Negative Interest Rate Policy

In recent years, a small but growing literature has started to argue that paper currency was never quite the obstacle to negative interest rates that it seemed. There are basically four approaches to implementing negative interest rates: 1) moving to a cashless society, since paying interest (positive or negative) on electronic bank reserves is no problem and already widespread practice; 2) finding a technological approach to paying interest (positive or negative) on paper currency, an idea that Keynes considered at length; 3) dispensing with the one-to-one exchange rate between electronic bank reserves and paper currency, which frees up the central bank to introduce approaches to discounting cash that mimic paying negative interest; and 4) taking steps to make large-scale hoarding of cash much more costly—for example, by phasing out large-denomination notes—without affecting normal retail cash transactions.

Curiously, to the limited extent the modern macroeconomics literature has discussed breaking the zero bound, options 1 and 2 have received virtually all the attention, even though neither is really viable. Eliminating cash would certainly obliterate the zero bound on interest rates because it is trivial to pay negative

interest on electronic money, unlike the situation with paper money. But for reasons of maintaining privacy, providing a safety valve to regulations, and offering a backup payment mechanism during internet/power outages, moving to a completely cashless society remains too high a price to pay simply to expand the central bank toolkit.

Directly paying negative interest rates on anonymous physical currency is also a nonstarter, though there have been some very creative suggestions for how it might be done. In *The General Theory*, Keynes (1936) has an extended discussion of the early writings of maverick economist Silvio Gesell (1916), who had proposed paying negative interest rates on paper currency by requiring that stamps be purchased and periodically affixed to the back of each note. Writing before the advent of electronic banking, Keynes ultimately rejected the idea because he believed that there was no simple and practical way to pay a negative interest rate on money without making it extremely illiquid. Goodfriend (2000) updates Gesell's idea by proposing that instead of requiring people to periodically get their currency stamped, the government can embed magnetic strips. It then records the time individual bills have been outside the banking system and charges a negative interest rate accordingly when the bills are re-deposited. Aside from the cost of the infrastructure required to implement this plan, it would be difficult for individuals and small proprietors to know how long any given bill has been outside the banking system (and therefore how much to discount it in retail transactions), again making currency relatively illiquid.

One wonders whether Keynes might have re-evaluated his position, and perhaps even restated his analysis of monetary and fiscal policy at the zero bound, had he been aware of the dual currency proposal of Robert Eisler (1932), which in recent times has been taken up by Buiter (2009) and by Agarwal and Kimball (2015).[6] The idea of one country having two different currencies with an exchange rate between them may seem implausible, but the basics are not difficult to explain.

The first step in setting up a dual currency system would be for the government to declare that the "real" currency is electronic bank reserves and that all government contracts, taxes, and payments are to be denominated in electronic dollars. As we have already noted, paying negative interest on electronic money or bank reserves is a nonissue.

Say then that the government wants to set a policy interest rate of negative 3 percent to combat a financial crisis. To stop a run into paper currency, it would simultaneously announce that the exchange rate on paper currency in terms of electronic bank reserves would depreciate at 3 percent per year. For example, after a year, the central bank would give only .97 electronic dollars for one paper dollar; after two years, it would give back only .94. What is ingenious about this proposal (compared to Gessell's stamped-money) is that all currency notes sell at the same discount. No one needs to know how long an individual note has been outside the banking sector; all that matters is the current exchange rate of paper money for

[6] In Rogoff (2017), I note that the 13th-century emperor Kublai Khan, grandson of Genghis Khan, also imposed an exchange rate between currency inside and outside the Mongol Treasury.

electronic money. Observe that it would have been perfectly possible to implement Eisler's (1932) approach in the 1930s when bank accounting was kept on paper books, albeit considerably more cumbersome in the absence of computers.

The dual currency system is elegant, but it does raise some issues of its own. One issue is that paper currency and electronic currency are not perfect substitutes (which is why the interest rate on bank money can deviate for long periods from cash), and so finding the correct path for the exchange rate is not quite as straightforward as the preceding example suggests. A further subtle but important point is that the Eisler (1932) approach only gets around the zero-bound constraint if the private sector follows the government's lead in converting all contracts to electronic currency. In most advanced countries, private agents are free to contract on whatever indexation scheme they prefer; this is not a condition that can be imposed by fiat. If the private sector does not convert to electronic currency, the zero bound would re-emerge since it still exists for paper currency. Finally, one must consider that after a period of negative interest rates, paper and electronic currency would no longer trade at par, which would be an inconvenience in normal times. Restoring par would require a period of paying positive interest rates on electronic reserves, which might potentially interfere with other monetary goals.

The fourth approach to implementing negative interest rates is perhaps the crudest, yet in some ways the simplest. This approach starts with the observation that the zero bound on interest rates is not literally zero, because it is costly to transport, store, and insure large quantities of cash. This is why several central banks (including Switzerland, Sweden, Denmark, Japan, and the Eurozone) have been able to set small negative rates (for example, –0.75 percent in Switzerland) without setting off a massive run into cash. No one quite knows the practical limits of just how low central banks can bring interest rates before creating a chaotic run. A plausible guess might be perhaps –1 or –2 percent; the exact number is sensitive to the length of time that negative interest rate policy is expected to persist, because hoarding imposes both fixed and variable costs. Banks can easily use their existing vaults to store some extra cash, but if they try to store billions extra, insurance companies will charge a nonlinear premium to compensate for the risk of very large losses. Banks would also have to pay the fixed shipping and insurance costs of transporting the cash, all the while not knowing how long the negative rate episode would last. Private hoarding companies would face the same problems, not to mention that new vaults take time to build.

There are several ways large-scale hoarding costs might be made even more prohibitive, short of the dual currency system, while still exempting small depositors. One place to start would be by phasing out large-denomination notes that hoarders would naturally use to economize on shipping and storage costs. In Rogoff (1998, 2017), I argue that independent of monetary policy considerations, there is a strong case for phasing out large-denomination paper currency notes, starting with large bills like the US$100, the 500 euro note (about $570 today) and the 1,000 Swiss franc note (worth about $1,000). The argument is that even as paper currency

is becoming increasingly less important in medium- and large-scale legal transactions, it remains important in facilitating wholesale criminal activity and tax evasion. Large-denomination notes make up a huge fraction of the value of outstanding currency, even though relatively few people use them: for example, the US$100 bill represents 81 percent of the US currency supply, while notes with denominations of $10 and below account for only about 3 percent.

Getting rid of $50s and $100s would already multiply the bulk and weight of storage cash compared to $10 bills by a factor of five or ten, and yet would have very little effect on ordinary retail transactions and the vast majority of people who do not rely on big notes for any of their cash activities. After all, $100,000 in $10 bills can still fit into an ordinary-size briefcase.

Restricting currency to small denominations should suffice to raise hoarding costs beyond any threshold where a wholesale run into cash by large-scale financial institutions like pension funds, insurance companies, and others is likely to be cost-effective in a plausible negative interest rate episode. This is particularly the case if regulators impose high standards of insurance on bulk cash hoarders, as well as on reinsurance companies that insure against theft and loss. It would be easy to take further steps, like charging a fee for redepositing large amounts of cash into the banking sector, in part to help defray the considerable handling costs that central banks otherwise provide for free.

The idea is to go to a less-cash society, not a cashless one. Although the poor do not rely much on large-denomination notes, any transition should nevertheless include provision for financial inclusion, such as free or highly subsidized checking accounts for low-income individuals, which could also be used to facilitate government transfer payments that are now made by check.

Among the largest economies, Japan is arguably the most natural candidate for an early transition to a less-cash economy, especially as it has floundered around the zero bound for so long. Also, after Switzerland, Japan has the highest per-capita cash issuance of any advanced economy, even though its physical currency is not held much outside of Japan. True, large notes are widely used by everyday people in Japan, but nevertheless a large share of the large-denomination notes appear to support various forms of tax evasion and criminal activity.

A far less obvious candidate is India, because like most emerging markets and developing economies, its financial infrastructure remains underdeveloped. Nevertheless, in November 2016, Indian Prime Minister Narendra Modi demonetized the country's two largest bills, the 500 and 1000 rupee notes (worth about $7.50 and $15 at the time), giving citizens only 50 days to make the exchange (as opposed to taking up to seven years as in Rogoff 2017). One problem the Indian government faced due to the rapidity of its move was that the central bank did not have on hand nearly a large enough supply of new notes to exchange for the old ones. While demonetization may still lead to long-run benefits in a country like India, where tax evasion is widespread (less than 2 percent of people pay taxes) and corruption is rife, the Indian experience reinforces the case for making any changes to the transaction system slowly over a period of years.

Paper currency is hardly the only constraint on negative interest rates (McAndrews 2015; Rogoff 2017). Another obvious constraint is that if central banks charge negative interest rates on bank reserves, it might be difficult for private banks to pass these costs on to depositors. In fact, early experience in Europe has shown that banks can quite easily pass on negative interest rates to wholesale customers, such as pension funds and insurance companies, but they are reluctant to do so for small depositors. This obstacle can be overcome by allowing an exclusion for small retail customers, where banks are compensated by the central bank (or treasury) so they do not lose anything, say for deposits up to $1,000 per individual. The objective of negative interest rate policy is to achieve macroeconomic stabilization, not to raise revenues. Cynics might say the power of a central bank to employ negative interest rates, once granted, is likely to be abused, but central banks already have ample tools to abuse holders of cash and bank deposits through inflation and financial repression.

There are other obstacles. For example, with positive interest rates, lenders receive interest payments from borrowers; with negative interest rates, tax laws need to be adjusted so that lenders who are making interest payments get a deduction. As another example, consider those people who significantly overpay their estimated taxes and then claim a refund as an indirect way to make a loan to the government; with negative rates, it will be necessary for the government to charge individuals interest on such "loans." None of these obstacles is particularly difficult to handle given sufficient time. Early experimenters with negative rates such as Switzerland, Sweden, and Denmark, have confronted such problems and generally found that they can be negotiated straightforwardly.

Of course, there are also psychological obstacles to nominal negative interest rates basically stemming from money illusion: people are already used to having inflation drive real interest rates (the nominal interest rate minus expected inflation) deeply negative. But having even slightly negative nominal interest rates is a relatively new phenomenon in the paper currency era. (Back in the days of coinage, sovereigns routinely called in coins in exchange for newer ones with lower silver content, which effectively gave a negative nominal interest rate on currency.) One presumes that if small deposits are excluded and the various frictions are dealt with, the psychological obstacles will disappear; after all, a large fraction of the world supply of government bonds is already paying negative interest rates today.

Once the zero bound is cleared away, would central bank policy moves into negative interest rate territory necessarily operate the same way as traditional monetary policy? In theory, yes—real interest rates are what matter and we have had deeply negative real interest rates in the past. Lower nominal rates would stimulate investment and consumption demand through the same channels as in standard new Keynesian models, which typically abstract from currency (except for incorporating the zero bound).[7] In models with richer institutional settings that incorporate

[7]Cochrane (forthcoming) argues that for certain kinds of expectations mechanisms, it is possible to construct models where lowering the interest rate lowers inflation. Garcia-Schmidt and Woodford (2015) give a critique.

both bank reserves and currency, financial institutions have strong incentives to lend reserves into the financial system in some way rather than to hoard them.

Financial institutions have lobbied strongly against the early experiments with negative interest rates in Europe and Japan, complaining that they impinge on profitability, though in fact banks in Sweden and the Nordic countries have fared reasonably well over this period. Over the long run, though, with adjustments such as providing small savers with government-subsidized zero-interest accounts, financial institutions should be able to pass on the remaining costs. Many of the challenges that financial firms face today actually stem from a long period of negative real interest rates, and it is far from clear that allowing for negative nominal rates would worsen the problem, particularly of course if the negative nominal rates were used to combat another severe financial crisis or extremely deep recession.

Relatedly, some have objected to negative interest rate policy because it might exacerbate financial instability. But if central banks had access to open-ended negative rate policy, they might well be able to move the economy more quickly out of deep recession, particularly after a financial crisis, rather than be stuck in slow growth with zero interest rates for a decade. If negative interest rate policy works, it should promote financial stability.

There are, of course, are a wide variety of potential objections, ranging from concerns about money illusion (people care more about nominal rates than real rates), distrust of monetary authorities (though an irresponsible central bank already can wreak havoc through inflation), to those who naively believe the world should try to restore the pre-war gold standard. It is beyond the scope of this paper to treat all of these; the reader is referred to Rogoff (2017) for a more detailed discussion. One objection that tends to be vastly overstated is that negative overnight interest rates constitute an unfair tax on savers and pension holders, but this is a very narrow perspective. First, the issue of the zero bound comes up in no small part because central banks have been so restrained with respect to inflation, unlike the high-inflation 1970s, a period that was really bad for savers. Second, in a deep recession, significant negative short-term rates will raise longer-term inflation expectations as well as accelerate growth in output and employment. Nominal interest rates on sufficiently long-term bonds should rise. Last but not least, low policy interest rates typically push up equity and housing prices. So the blanket statement that negative rates are unambiguously bad for savers and pension holders is naive.

In principle, restoring the effectiveness of interest rate policy by fully removing the zero bound will make it possible to have central banks return to being limited-purpose institutions whose objective is to stabilize inflation and output. It might reduce pressures on them to take on large balance sheets and engage in directed credit and even fiscal stimulus. Over the long run, limiting the scope of central banks should help them maintain their independence.

What about the international implications? By making monetary policy more coherent and predictable, negative interest rate policy should help provide a more stable global capital market environment for emerging economies, as well as

provide a better basis for communication and cooperation among advanced-country central banks. Greater instrument transparency should alleviate the concerns about competitive depreciations and currency wars. Of course, there is much research to be done in understanding the subtleties of negative interest rate policy, but the objections once raised by Keynes back in the 1930s should no longer be considered definitive today.

Again, it is important to re-emphasize that if the road can be paved for effective negative interest rate policy (in contrast to the early experiments in Europe and Japan where cash has not been dealt with and many other frictions remain), then episodes would presumably be much shorter-lived than today's zero-bound episode, since monetary policy would not be paralyzed in reflating the economy. The main goal of enabling negative nominal interest rate policy is as a tool for dealing with very deep recessions, not as a routine policy. In normal times, central banks that want to debase currency already have ample tools by using inflation.

Conclusion

The international monetary system stands at a crossroads. Central banks, the linchpins of the global financial system, have come under enormous pressure in recent years as the zero bound on interest rates has forced them to employ alternative instruments. These alternative methods of conducting monetary policy expand the remit of central banks far beyond the limited-purpose institutions they had become in the era of financial market liberalization, and risk subjecting them to greater political interference and even loss of independence. In addition, there are significant theoretical and empirical questions about how well these alternative monetary instruments really work. The zero bound has confounded domestic macroeconomic policy and made international monetary policy extremely difficult, with countries accusing each other of trying to manipulate exchange rates in lieu of being able to affect interest rates. In the long run, undertaking institutional changes that clear the way for effective negative interest rate policy is likely to be the cleanest approach to restoring the efficacy of monetary policy at the zero bound. Creating the preconditions for effective negative interest rate policy will certainly require a number of tax, legal, and institutional changes in addition to dealing with cash, but the early experiences in Europe and Japan suggest these are manageable.

Eliminating the zero bound will not make an aging economy young, nor will it transform an economy with low productivity growth into a powerhouse of innovation. But effective negative interest rate policy can help monetary authorities in fighting deep recessions. In addition, it should end discussion of third-best alternatives to monetary policy facing a zero bound such as indefinitely postponing structural reform, reneging on trade agreements, and using fiscal policy to an extent beyond what normal cost–benefit analysis would suggest. Enabling effective negative rate policy is also much cleaner and more elegant than the second-best policy

of raising inflation targets. It thus could have many benefits in helping to foster a smoother and more natural functioning of the global financial system.

In an era where cash is becoming less important in the legal tax-compliant economy outside small-scale transactions, and where banking and retail transactions are increasingly electronic, it is perhaps time for macroeconomists to stop treating the zero bound as an immutable constant of nature. The zero lower bound was a major problem in the 1930s and again in the most recent global financial crisis. It does not need to be a major obstacle in the next one, and there are perfectly viable ideas for eventually solving it without going all the way to a cashless economy.

References

Agarwal, Ruchir, and Miles Kimball. 2015. "Breaking through the Zero Lower Bound." International Monetary Fund Working Paper 15/224.

Ascari, Guido, and Argia M. Sbordone. 2014. "The Macroeconomics of Trend Inflation." *Journal of Economic Literature* 52(3): 679–739.

Barro, Robert J. 2006. "Rare Disasters and Asset Markets in the Twentieth Century." *Quarterly Journal of Economics* 121(3): 823–66.

Bernanke, Ben S. 2002. "Deflation: Making Sure 'It' Doesn't Happen Here." Remarks before the National Economists Club, Washington, DC, November 21. https://www.federalreserve.gov/boarddocs/speeches/2002/20021121/default.htm.

Bernanke, Ben S. 2005. "The Global Savings Glut and the U.S. Current Account Deficit." Remarks at the Sandridge Lecture, Virginia Association of Economists, Richmond, Virginia, March 10. https://www.federalreserve.gov/boarddocs/speeches/2005/200503102/.

Blanchard, Olivier, Giovanni Dell'Ariccia, and Paolo Mauro. 2010. "Rethinking Macroeconomic Policy." *Journal of Money, Credit and Banking* 42(Issue Supplement 1): 199–215.

Bodenstein, Martin, Luca Guerrieri, and Christopher J. Gust. 2013. "Oil Shocks and the Zero Bound on Nominal Interest Rates." *Journal of International Money and Finance* 32(1): 941–67.

Buiter, Willem H. 2009. "Negative Nominal Interest Rates: Three Ways to Overcome the Zero Lower Bound." NBER Working Paper 15118.

Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2016. "Global Imbalances and Currency Wars at the ZLB." Unpublished paper.

Canzoneri, Matthew B., Dale W. Henderson, and Kenneth S. Rogoff. 1983. "The Information Content of the Interest Rate and Optimal Monetary Policy." *Quarterly Journal of Economics* 98(4): 545–66.

Carvalho, Carlos, Andrea Ferrero, and Fernanda Nechio. 2016. "Demographics and Real Interest Rates: Inspecting the Mechanism." Federal Reserve Bank of San Francisco Working Paper 2016–5.

Cecchetti, Stephen, and Kim Schoenhotlz. 2016. "A Primer on Helicopter Money." *Vox*, August 19. http://voxeu.org/article/primer-helicopter-money.

Christiano, Lawrence, Martin Eichenbaum, and Sergio Rebelo. 2011. "When Is the Government Spending Multiplier Large?" *Journal of Political Economy* 119(1): 78–121.

Cochrane, John. Forthcoming. "Michelson-Morley, Occam and Fisher: The Radical Implications of Stable Inflation at Near-Zero Interest Rates." In *NBER Macroeconomics Annual 2017, Volume 32,* edited by Martin Eichenbaum and Jonathan A. Parker. National Bureau of Economic Research.

Delong, J. Bradford, and Lawrence H. Summers. 2012. "Fiscal Policy in a Depressed Economy." *Brookings Papers on Economic Activity* 1: 233–97.

Eggertsson, Gauti. 2010. "The Paradox of Toil." Federal Reserve Bank of New York Staff Report

433.

Eggertsson, Gauti, Andrea Ferrero, and Andrea Raffo. 2014. "Can Structural Reforms Help Europe?" *Journal of Monetary Economics* 61: 2–22.

Eggertsson, Gauti B., Neil R. Mehrotra, Sanjay R. Singh, and Lawrence H. Summers. 2016. "A Contagious Malady? Open Economy Dimensions of Secular Stagnation." NBER Working Paper 22299.

Eichengreen, Barry. 1986. "The Political Economy of the Smoot–Hawley Tariff." NBER Working Paper 2001.

Eichengreen, Barry. 2016. "What's the Problem with Protectionism?" *Project Syndicate*, July 13.

Eisler, Robert. 1932. *Stable Money: The Remedy for the Economic World Crisis. A Programme of Financial Reconstruction for the International Conference, 1933.* With a Preface by Vincent C. Vickers. Search Publishing.

Farhi, Emmanuel, and Matteo Maggiori. 2016. "A Model of the International Monetary System." NBER Working Paper 22295.

Fuhrer, Jeffrey C., and Brian F. Madigan. 1997. "Monetary Policy When Interest Rates Are Bounded at Zero." *Review of Economics and Statistics* 79(4): 573–85.

Garcia-Schmidt, Mariana, and Michael Woodford. 2015. "Are Low Interest Rates Deflationary? A Paradox of Perfect-Foresight Analysis." NBER Working Paper 21614.

Geanakoplos, John. 2014. "Leverage, Default, and Forgiveness: Lessons from the American and European Crises." *Journal of Macroeconomics* 39(Part B): 313–33.

Gesell, Silvio. 1916 [1958]. *Die Natuerliche Wirtschaftsordnung [The Natural Economic Order].* London: Peter Owen, 1958.

Goodfriend, Marvin. 2000. "Overcoming the Zero Bound on Interest Rate Policy." *Journal of Money, Credit, and Banking* 32(4): 1007–35.

Gordon, Robert J. 2016. *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War.* Princeton University Press.

Gourinchas, Pierre-Olivier, and Helene Rey. 2016. "Real Interest Rates, Imbalances and the Curse of Regional Safe Asset Providers at the Zero Lower Bound." Paper presented at the European Central Bank Forum on Central Banking, Sinatra, Portugal.

Greenwood, Robin, Samuel G. Hanson, and Jeremy C. Stein. 2016. "The Federal Reserve's Balance Sheet as a Financial-Stability Tool." Paper presented at the Federal Reserve Bank of Kansas City Symposium on Economic Policy, Jackson Hole, WY.

Henry, James. 1976. "Calling in the Big Bills." *Washington Monthly*, May, 22–33.

Holston, Kathryn, Thomas Laubach, and John C. Williams. 2016. "Measuring the Natural Rate of Interest: International Trends and Determinants." Federal Reserve Bank of San Francisco Working Paper 2016–11.

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

Keynes, John Maynard. 1936. *The General Theory of Employment, Interest, and Money.* Macmillan.

Kozlowski, Julian, Laura Veldkamp and Venky Venkateswaran. 2017. "The Tail That Wags the Economy: Beliefs and Persistent Stagnation." http://people.stern.nyu.edu/lveldkam/pdfs/KVVTailEvents.pdf.

Krugman, Paul R. 1998. "It's Baaack: Japan's Slump and the Return of the Liquidity Trap." *Brookings Papers on Economic Activity* 2: 137–205.

Laubach, Thomas, and John C. Williams. 2015. "Measuring the Natural Rate of Interest Redux." Federal Reserve Bank of San Francisco Working Paper Series 2015–16.

Lebow, David E. 1993. "Monetary Policy at Near-Zero Interest Rates." Board of Governors of the Federal Reserve System Working Paper, Economic Activity Section, no 136.

McAndrews, James. 2015. "Negative Nominal Central Bank Policy Rates: Where Is the Lower Bound?" Remarks at the University of Wisconsin, Madison, Wisconsin, May 8. https://www.newyorkfed.org/newsevents/speeches/2015/mca150508.html.

Nakamura, Emi, Jon Steinsson, Patrick Sun, and Daniel Villar. 2016. "The Elusive Costs of Inflation: Price Dispersion during the U.S. Great Inflation." NBER Working Paper 22505.

Ohanian, Lee E. 2001. "Why Did Productivity Fall So Much during the Great Depression?" *American Economic Review* 91(2): 34–38.

Rajan, Raghuram. 2016. "New Rules for the Monetary Game." *Project Syndicate*, March 21.

Reifschneider, David L. 2016. "Gauging the Ability of the FOMC to Respond to Future Recessions." Finance and Economics Discussion Series, no. 2016-068, Board of Governors of the Federal Reserve System.

Reinhart, Carmen M., Vincent Reinhart, and Kenneth Rogoff. 2015. "Dealing with Debt." *Journal of International Economics* 96(Supplement 1): S43–S55.

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

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

Rogoff, Kenneth S. 1998. "Blessing or Curse? Foreign and Underground Demand for Euro Notes." *Economic Policy* 13(26): 261–303.

Rogoff, Kenneth S. 2015. "Costs and Benefits to Phasing out Paper Currency." In *National Bureau of Economic Research Macroeconomics Annual 2014,* vol. 29, edited by Jonathan A. Parker and Michael Woodford, 445–56. University of Chicago Press.

Rogoff, Kenneth S. 2016. "Debt Supercycle, Not Secular Stagnation." In *Progress and Confusion: The State of Macroeconomic Policy,* edited by Olivier J. Blanchard, Raghuram G. Rajan, Kenneth S. Rogoff, and Lawrence H. Summers, 19–28. MIT Press.

Rogoff, Kenneth S. 2017. *The Curse of Cash: How Large-Denomination Bills Aid Tax Evasion and Crime and Constrain Monetary Policy.* Paperback edition. Princeton University Press.

Summers, Lawrence. 2013. "Secular Stagnation." *Business Economics* 49(2): 65–73.

Taylor, John B. 1993. "Discretion versus Policy Rules in Practice." *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.

Werning, Ivan. 2011. "Managing a Liquidity Trap: Monetary and Fiscal Policy." NBER Working Paper 17344.

Woodford, Michael. 2012. "Methods of Monetary Accommodation at the Interest-Rate Lower Bound." Proceedings of the Federal Reserve Bank of Kansas City Symposium on Economic Policy, Jackson Hole, Wyoming, August 30–September 1. https://www.kansascityfed.org/publications/research/escp/symposiums/escp-2012.

Wu, Jing Cynthia, and Fan Dora Xia. 2016. "Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound." *Journal of Money, Credit, and Banking* 48(2–3): 253–91.

Yellen, Janet L. 2016. "The Federal Reserve's Monetary Policy Toolkit: Past, Present, and Future." Speech at the Federal Reserve Bank of Kansas City Symposium on "Designing Resilient Monetary Policy Frameworks for the Future," Jackson Hole, Wyoming, August 26. https://www.federalreserve.gov/newsevents/speech/yellen20160826a.htm.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 67–88

Is the US Public Corporation in Trouble?

Kathleen M. Kahle and René M. Stulz

In his famous 1989 *Harvard Business Review* article titled "Eclipse of the Public Corporation," Jensen (1989) predicted the demise of the public corporation. He argued that public corporations are inefficient organizational forms because private firms financed by debt and private equity can resolve agency conflicts between investors and managers better than public firms. His prediction initially appeared invalid. The number of public firms increased sharply in the first half of the 1990s. However, the number of listed firms peaked in 1997 and has since fallen by half, such that there are fewer public corporations today than 40 years ago (Doidge, Karolyi, and Stulz 2017). Does this fall vindicate Jensen's (1989) argument? Is the public corporation in trouble?

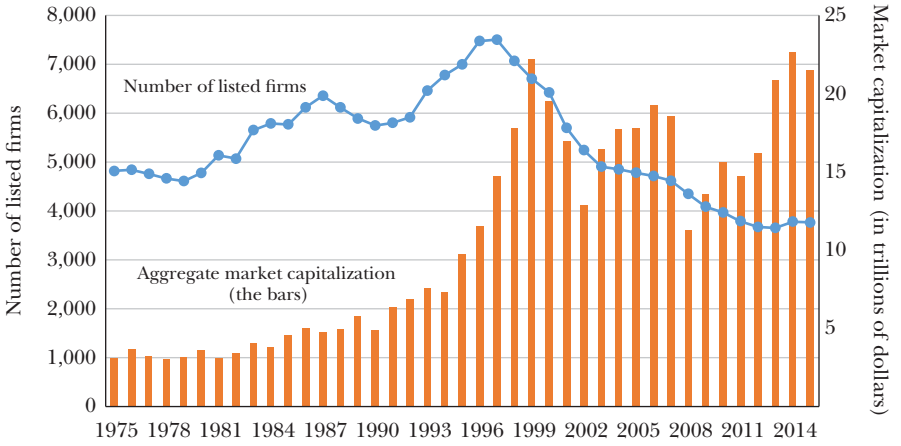
In this paper, we examine the evolution of US public corporations over the last 40 years. Over this time period, the universe of US public corporations experienced massive changes. Not only are there fewer public corporations today than 40 years ago, but these corporations are very different. They are older and larger. They are in different industries. Their asset structure has changed, as they invest less in physical assets, but more in R&D. They finance themselves differently. They are less profitable on average, but profitability increases with size, so total profits of US public corporations are higher. Total payouts to shareholders are higher, but these

■ *Kathleen Kahle is the Thomas C. Moses Professor in Finance, Eller College of Management, University of Arizona, Tucson, Arizona. René Stulz is the Everett D. Reese Chair of Banking and Monetary Economics, Fisher College of Business, Ohio State University, Columbus, Ohio, and is a Research Associate at the National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are kkahle@eller.arizona.edu and stulz@fisher.osu.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.67

doi=10.1257/jep.31.3.67

Figure 1
Number of Listed Firms by Year on the NYSE, Nasdaq, and Amex, and Market Capitalization from 1975 to 2015



Source: The source for number of listings and market capitalization is Center for Research in Security Prices (CRSP) data.
Note: The market capitalization is shown in 2015 dollars.

payouts now are often in the form of share repurchases rather than dividends. Their shareholders are very different, as institutions now typically hold more than half the shares of large corporations.

To illustrate how US public corporations have changed, we compare snapshots in 1975, 1995, and 2015. The variables we discuss are reported in Table 1 for these three years. These three snapshots correspond to the beginning and the end of our sample period, as well as a year in the middle, which is close to the peak in the number of public corporations. In the following sections, we discuss each section of Table 1: patterns in the number and age of listed firms, valuation, investment, profitability, financing, ownership, and payout policy. We conclude with some thoughts about the meaning of these patterns for public firms in the United States.

The Number and Age of Public Firms

Figure 1 shows the evolution of the number of listings of US firms from 1975 to 2015, including firms listed on the New York Stock Exchange (NYSE), Amex, and Nasdaq. In 1975, the US economy has 4,819 listed firms, as also shown in Table 1.[1]

[1]We use two main data sources for our analysis: Center for Research in Security Prices (CRSP) and Compustat. From CRSP we obtain all US firms (share codes 10 and 11) listed on the NYSE, Amex, and Nasdaq, excluding investment funds and trusts (Standard Industrial Classification (SIC) codes 6722,

Table 1
Mean Characteristics

	1975	*1995*	*2015*	*t*-test *75 vs. 95*	*t*-test *95 vs. 15*	*t*-test *75 vs. 15*
Number of listed firms	4,819	7,002	3,766			
Age	10.9	12.2	18.4	***	***	***
Valuation						
Market Cap/GDP	38.4%	78.0%	116.2%	***	***	***
Tobin's *q*	0.769	1.731	1.639	***	**	***
Market cap (millions of dollars)	662.8	1,400.1	5,752.9	***	***	***
Small firms	61.5%	43.9%	22.6%	***	***	***
Revenue Herfindahl	1,391.5	811.7	1,179.5	***	***	***
Investment						
Capital expenditures/Assets	8.0%	9.6%	4.2%	***	***	***
R&D/Assets	1.3%	5.7%	7.5%	***	***	***
Fixed assets/Assets	34.7%	25.4%	19.7%	***	***	***
Inventory/Assets	23.6%	12.9%	8.2%	***	***	***
Cash/Assets	9.2%	15.6%	21.6%	***	***	***
Profitability						
Operating cash flow/Assets	8.5%	2.9%	–4.2%	***	***	***
Loss firms	13.6%	29.4%	37.2%	***	***	***
R&D-adjusted operating cash flow/Assets	9.8%	8.6%	3.3%	***	***	***
Return on assets (ROA)	4.3%	–3.3%	–8.3%	***	***	***
Financing						
Book leverage	26.6%	21.0%	22.7%	***	***	***
Market leverage	28.5%	15.5%	15.8%	***		***
Net leverage	17.4%	5.4%	1.3%	***	***	***
Negative net leverage firms	23.7%	39.7%	43.1%	***	***	***
Interest/Assets	2.6%	2.7%	1.8%	**	***	***
No debt firms	6.1%	12.7%	17.3%	***	***	***
Net equity issuance	0.5%	25.2%	15.4%	***	***	***
Ownership						
Institutional ownership[a]	17.7%	29.8%	50.4%	***	***	***
Blockholder[a]	11.9%	19.5%	32.0%	***	***	***
Payout policy						
Dividend paying firms	63.5%	34.0%	41.9%	***	***	***
Dividends/Assets	1.3%	0.7%	1.0%	***	***	***
Repurchases/Assets	0.3%	0.6%	2.0%	***	***	***
Total payout/Assets	1.6%	1.4%	3.2%	***	***	***
Total payout/Net income	27.1%	20.5%	47.0%	***	***	***

Note: Detailed descriptions of the variables are provided in the online Appendix at the journal website.
[a] Data is not available in 1975 so we use values for the first year data is available.
***, **, and * indicate significance at the 1 percent, 5 percent, and 10 percent levels, respectively.

6726, 6798, and 6799). When examining Compustat data, we use the intersection of CRSP and Compustat firms. As for firms listed on CRSP that are not covered by Compustat, we find that these firms account for 1–3 percent of the aggregate market capitalization of all listed firms.

This number increases rather steadily until 1997, when it reaches 7,507 listed firms. After that, the number falls rapidly until 2003 and then continues to fall at a slower pace, before leveling out around 2013. There are 3,766 listed firms in 2015, a number that is over 20 percent (1,053 firms) lower than 40 years before. In 1975, the US economy has 22.4 publicly listed firms per million inhabitants. In 2015, it has just 11.7 listed firms per million inhabitants.

As a result of the decrease in the number of listed firms, the US economy has developed a "listing gap" in that it has fewer listed firms than expected (Doidge, Karolyi, and Stulz 2017). Specifically, if the variables that explain the number of listings per capita worldwide—like dimensions of economic development and institutions—are used to predict the number of listed firms in the United States, the prediction is roughly equal to the actual number prior to 1999; by 2012, however, the predicted number is more than double the actual number. In short, there is no listing gap in 1998, but a gap emerges after this.

The steady decrease in the number of listed firms since 1997 results from both low numbers of newly listed firms and high numbers of delists. The majority of new lists are due to initial public offerings. However, the number of initial public offerings decreases dramatically after 2000, such that the average yearly number of initial public offerings after 2000 is roughly one-third of the average from 1980 to 2000 (Gao, Ritter, and Zhu 2013; Doidge, Karolyi, and Stulz 2013).

The three main reasons for a public firm to delist are: 1) it no longer meets the listing requirements, which is typically due to financial distress, 2) it has been acquired, or 3) it voluntarily delists. Doidge, Karolyi, and Stulz (2017) find that mergers are the dominant reason for delisting since the listing peak in 1997. Firms that voluntarily delist can either keep trading over-the-counter or become private firms. The contribution of firms that voluntarily delist to the number of delists is small compared to the contribution of acquisitions.

We also examine the evolution of firm age. There are two ways to measure the age of a firm: from the date of incorporation or from the date the firm went public. Hathaway and Litan (2014) study the age since incorporation for all US firms, both private and public. They conclude that the increase in the older firms' share of economic activity is "a trend that has occurred in every state and metropolitan area, in every firm size category, and in each broad industrial sector." This aging trend is more dramatic among public firms than private firms. Unfortunately, it is difficult to assess the age since incorporation of public firms because public databases lack systematic information on the age of incorporation. Data for the age since listing is available, but this data has an important limitation too. Nasdaq firms were added to existing databases at the beginning of the 1970s and were given a listed age of zero when they were added, even though these firms were already public. As a result, the average age since listing of 10.9 years in 1975 reported in Table 1 is biased downward. Despite this bias, the average age changes little over the next 20 years. In 1995, average age is 12.2 years. The reason for the relative stability of age from 1975 to 1995 is that the number of public firms increases, so the increase in age of the older firms is offset by the influx of young firms. However, from 1995 to 2015, the age of

public firms increases to 18.4 years. Median age also increases in the last 20 years. The median age is 8 years in 1995, 6.3 years in 1997, and since then the median age has increased by a factor of 2.5. The aging of US public firms has implications for how these firms behave: for instance, Loderer, Stulz, and Walchli (forthcoming) find that older firms innovate less and are more rigid.

Valuation and Concentration of Public Firms

The aggregate market capitalization of listed firms in 2015—the sum of the market value of all listed firms—is about seven times higher than in 1975 (expressed in 2015 dollars). However, aggregate market capitalization does not evolve smoothly. In particular, between 1999 and 2015, the aggregate market capitalization of listed firms experiences two sharp drops. As illustrated by the bars in Figure 1, the aggregate market capitalization changes from about $22 trillion at the peak of the dot-com bubble in 1999 to $11 trillion in 2008 and then back to about $22 trillion by 2015.

Many academic studies compare the aggregate market capitalization of stocks to GDP as a measure of financial development (as discussed in Levine 1997). Table 1 shows that this ratio is higher in 2015 than either in 1995 or in 1975, but like market capitalization, this ratio is volatile. It is 38.4 percent in 1975, climbs to 78.0 percent in 1995, peaks at 153.5 percent in 1999, drops to 69.2 percent in 2008, and rises back to 116.2 percent in 2015. The ratio is 24 percent lower in 2015 than at its peak in 1999.

An often-used valuation ratio for firms is Tobin's *q*, the ratio of the market value of the firm's assets to the replacement cost of the assets. Using the market value of assets divided by the book value of assets as a proxy for Tobin's *q*, as is commonly done in corporate finance,[2] Tobin's *q* is 2.14 at the peak of the dot-com bubble in 1999. In contrast, it is 0.77 in 1975, 1.73 in 1995, and 1.64 in 2015.

Whether we examine average or median firm market capitalization, firms have become larger since 1975. We first measure the average size of listed firms using market capitalization, again expressed in 2015 dollars. In 1975, the mean market capitalization is a bit more than one-tenth the mean market capitalization in 2015: $663 million versus $5,753 million, as shown in Table 1. A similar evolution takes place for the median market capitalization (not tabulated here), which increases from $60 million to $570 million. Mean market capitalization increases by 299 percent in the 22 years before the 1997 peak in new listings, and then increases by 290 percent in the 18 years since 1997.

The distribution of market capitalization is extremely skewed, although the level of skewness is similar in 1975 and 2015, with a large increase in skewness in the late 1990s. The ratio of mean to median is 11.0 in 1975 and 10.1 in 2015, but peaks

[2] To obtain the market value of assets, the practice in corporate finance is essentially to replace the book value of equity with its market value.

at 21.4 in 2000. Another way to analyze the distribution of market capitalization is to look at the smallest and the largest number of firms it takes to reach 25 percent of the market's total capitalization. In 1975, the 14 largest firms have an aggregate market capitalization equal to 25 percent of the market as a whole, as do the 4,484 smallest firms, or 93.0 percent of all listed firms. In 2015, the 21 largest firms have a total market capitalization equal to 25 percent of the market as a whole, as do the 3,487 smallest firms (92.6 percent of listed firms).

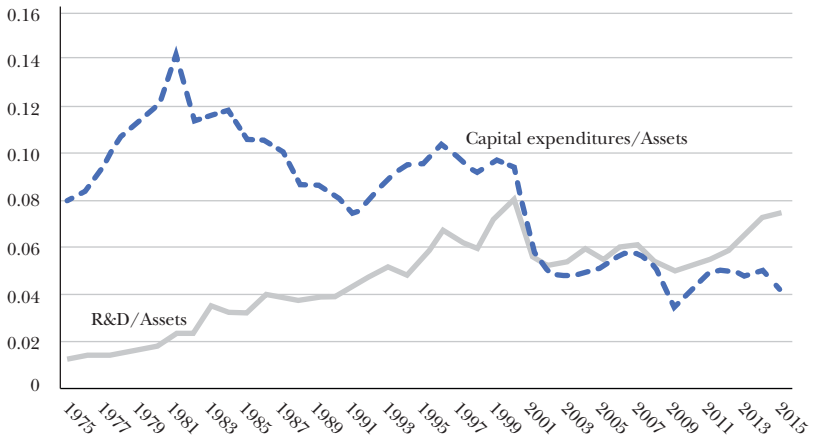
In short, while listed firms are larger today than 40 years ago in terms of market capitalization, the distribution of firm size in 2015 is similar to 1975—with both being more concentrated than the distribution in 1995. These patterns have given rise to concerns about whether markets have become less receptive to small firms.[3] A simple but rough benchmark is to compute the percentage of listed firms that are small, defined as having a market capitalization of less than $100 million in 2015 dollars. In 1975, 61.5 percent of listed firms are small, as shown in Table 1. This percentage peaks at 63.2 percent in 1990, and then falls. The share of small, listed firms dropped all the way to 19.1 percent of listed firms in 2013, before rebounding slightly to 22.6 percent in 2015. In other words, small listed firms are much scarcer today than 20 or 40 years ago.

One obvious concern with fewer but larger firms is that concentration within industries can increase, which could possibly adversely affect competition. To examine this, we construct a Herfindahl index of revenue at the three-digit NAICS (North American Industry Classification System) level for public firms.[4] We find that the average Herfindahl index at the industry level increases by 45 percent from 1995 to 2015, and from 811.7 to 1,179.5, as shown in Table 1. However, the average index is significantly lower in 2015 compared to 1975, when it is 1,391.5. In other words, three-digit NAIC industries are on average much more concentrated now than 20 years ago, but less than 40 years ago. An obvious limitation of this analysis is that it ignores foreign firms, whose importance has increased substantially over the past 40 years, and also private firms. Hence, the increase in Herfindahl ratios since 1995 may overstate the potential increase in concentration. Grullon, Larkin, and Michaely (2016) take private firms into account in studying the increase of industry-level concentration in the US economy and find this does not change conclusions about the increase in concentration. Though an increase in concentration could lead to a decrease in competition, of course this is not necessarily the case.

[3]For instance, Weild and Kim (2010) argue that market structure has decreased the benefits of listing for small firms, and Gao, Ritter, and Zhu (2013) propose that growing economies of scope make it more advantageous for firms to be acquired by larger firms before an initial public offering.

[4]A Herfindahl index is constructed by taking the market share of each firm in an industry, squaring it, and then summing to a total. Thus, an industry ruled by a monopoly with 100 percent of the market will have a Herfindahl of 10,000 (that is, 100^2), while an industry with 100 firms that each have 1 percent of the market will have a Herfindahl of 100 (that is, 100×1^2).

Figure 2
Evolution of Capital Expenditures and R&D from 1975 through 2015
(as a ratio of total assets)



Source: The sample is composed of listed firms on CRSP (Center for Research in Security Prices) for which Compustat data are available. Accounting data are from Compustat.
Note: Detailed variable definitions are in the online Appendix at the journal website.

Investment

From 1975 to 2015, research and development (R&D) investment and, more generally intangible assets, became increasingly important for the production of goods in the US economy, which has implications for how firms invest, perform, and finance themselves. This is reflected in the path of various types of investment, as shown in Figure 2 and Table 1. First we consider the evolution of capital expenditures over time. The US economy is relatively weak in 1975, so it is not surprising that the average ratio of capital expenditures to assets increases at first, peaking in 1981 at 14.1 percent. By 1988, the ratio of capital expenditures to assets falls below 10 percent, and after rebounding to 10.5 percent in 1996, the ratio drops and averages 4.5 percent from 2009 to 2015. It is noteworthy that average capital expenditures as a fraction of assets in 2015 are less than in 2008, the year of the financial crisis.[5]

The increase in the importance of intangible assets can also be seen by examining the largest firms over time. In 1975, the largest firm by market capitalization is IBM. Besides IBM, the other firms in the top five are AT&T, Exxon, Eastman

[5] In results not tabulated here, the same evolution takes place if we use an asset-weighted average instead of an equally weighted average. In this case, capital expenditures are 9.8 percent of assets in 1975, 5.1 percent in 1995, and 2.6 percent in 2015. Strikingly, the asset-weighted average of capital expenditures drops below 3 percent in 2002 and has not exceeded 3 percent since then.

Kodak, and General Motors. Exxon is the only firm that remains in the top five in all three of our snapshot years. In 2015, the top five firms by market capitalization are, starting from the largest, Apple, Google, Microsoft, ExxonMobil (after the merger in 1999), and Amazon. In 1975, the average ratio of capital expenditures to assets for the top five firms is 13 percent, while the average ratio of R&D expenditures to assets is 4 percent. By 2015, the capital expenditures ratio drops to 6 percent while the research and development ratio increases to 9 percent.

This change in the relative importance of R&D versus capital expenditures for the five largest firms has taken place across listed firms as a whole. Listed firms have a much lower average ratio of capital expenditures to assets and a much higher ratio of R&D expenditures to assets in 2015 than they do in 1975. Figure 2 shows the evolution of average R&D to assets over time. As reported in Table 1, the equally weighted average of R&D to assets is 1.3 percent in 1975, 5.7 percent in 1995, and 7.5 percent in 2015. Around 2001, R&D expenditures start slightly exceeding capital expenditures, and the gap grows in recent years. In 2015, R&D expenditures by listed firms are 78 percent higher than capital expenditures.[6] Overall, the rise in R&D expenditures does not offset the decrease in capital investment. If we sum R&D and capital expenditures as a measure of total investment, its lowest value during our sample period is 8.5 percent in 2009. Total investment peaks at 17.5 percent in 2000. In 2015, it is only 11.6 percent, but it does not exceed 12 percent after 2000.

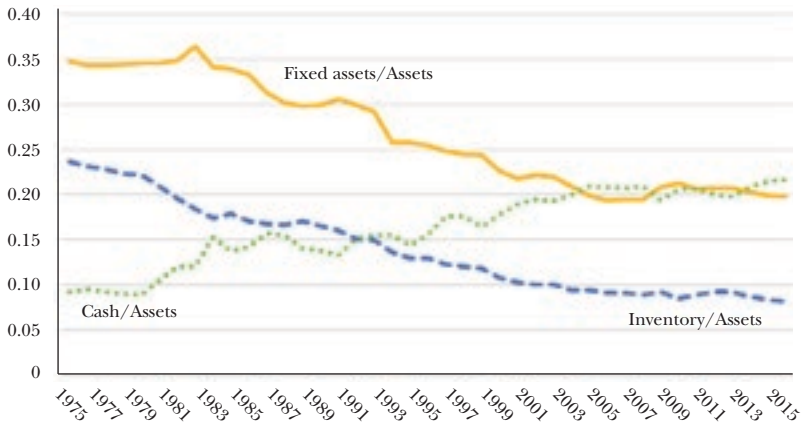
Given the decline in capital investment, it is not surprising that listed firms have experienced a decrease in the fraction of assets that are "fixed assets"—that is, property, plant, and equipment. Figure 3 shows the evolution of the equally weighted ratio of fixed assets to total assets. In 1975, the equally weighted average is 34.7 percent (as shown also in Table 1). By 2015, it is 19.7 percent. While publicly available databases do not make it possible to assess the extent to which firms substitute outsourcing for in-house production, these results are consistent with an increase in outsourcing, which increases substantially over our sample period (da Silveira 2014).

Inventory holdings also fall dramatically over our time period, as shown in Figure 3, partly due to the introduction of just-in-time production processes in which firms receive goods only when needed. As reported in Table 1, the equally weighted ratio of inventories to assets is 23.6 percent of assets in 1975. By 2015, that ratio is just 8.2 percent.

Though public firms today have lower levels of fixed assets and inventories, they hold more cash. As shown in Figure 3 and Table 1, the equally weighted ratio of cash to assets is 9.2 percent in 1975, and more than doubles to 21.6 percent in 2015. The increase in cash holdings is not as noticeable for large firms, but the average ratio of cash to assets for the five largest firms by market capitalization is 23 percent in 2015; these firms hold $243 billion in cash. In contrast to the equally weighted

[6] This shift in how firms invest is fairly dramatic when we examine averages, but not as large when we look at medians. A primary reason for this difference is that the median firm does not report *any* research and development expense.

Figure 3
Evolution of Fixed Assets, Inventory, and Cash Holdings
(as a ratio of total assets)



Source: The sample is composed of listed firms on CRSP (Center for Research in Security Prices) for which Compustat data are available. Accounting data are from Compustat.
Note: Detailed variable definitions are in the online Appendix at the journal website.

average, the asset-weighted average of cash to assets (not tabulated separately here) falls in the 1980s, reaching a low of 7.9 percent in 1990. The ratio then increases and peaks at 13.3 percent in 2013; it is 12.6 percent in 2015. It is well-documented that firms with more intangible assets and more R&D expenditures hold more cash and that the increase in R&D expenditures helps explain the increase in cash holdings (Bates, Kahle, and Stulz 2009).

One concern about measures of investment is that investments in most intangible assets—like organizational capital or benefits from accumulated past R&D investments—are not recorded on firms' balance sheets. Accounting rules dictate that investments in intangible assets are expensed, even though the importance of these assets seems to be rising over time. To the extent that intangible assets become more important over the period we consider, a firm's balance sheet becomes a less-informative measure of the firm's financial position. Eisfeldt and Papanikolaou (2014) define organizational capital as the intangible capital that relies on human inputs, including the firm-specific human capital of employees that enables firms to work more efficiently. Estimates of the importance of intangible assets for US firms vary. Falato, Kadyrzhanova, and Sim (2013) find that intangible capital averaged 10 percent of net assets (assets minus cash holdings) in 1970, slightly higher in 1975, and then increased steadily to exceed 50 percent in 2010. They also find that capitalized R&D represents about one-third of intangible capital and organizational capital roughly two-thirds. Corrado, Hulten, and Sichel (2009) argue that organizational capital is the largest component of intangible capital, and accounts

for about 30 percent of all intangible assets. Eisfeldt and Papanikolaou (2014) show that organizational capital is more important than investment in property, plant, and equipment in the health, high tech, and finance industries but less important in manufacturing and consumer industries. For finance, high tech, and health industries, the ratio of organizational capital to property, plant, and equipment increases steadily since 1995 and is at or close to a peak in 2012 (the end of their sample period).

Profitability

One well-accepted measure of profitability is the ratio of a firm's operating cash flow to assets. We define operating cash flow as operating income before depreciation minus interest and taxes; assets are measured at the beginning of each time period. As shown in Table 1, the equally weighted average of this ratio across listed firms falls sharply during our sample period. It averages 4.3 percent from 1975 until 1995, and 0.2 percent since 1995. Surprisingly, this measure of cash flow is never negative before 1998; since then, it is negative in seven years, including the last three years of our sample.

If we asset-weight rather than equal-weight the operating cash flow measures, average cash flow and average adjusted cash flow are higher, which indicates that larger firms have a higher ratio of cash flow to assets. Another way to see this is by separating the firms in the top decile of assets from the firms in all the other deciles. The equally weighted average of cash flow to assets is marginally higher after 1995 compared to before (8.3 percent versus 8.2 percent) for the largest firms. Average cash flow for the largest firms is never negative and its minimum value is 6.7 percent in 1982. In contrast, the equally weighted cash flow for the other firms is negative only once before 1995, but after 1995, it is negative 11 times. Therefore, firms have been performing poorly on average, except for the largest firms. Further evidence of poor performance can be found in the fact that the fraction of firms with negative net income increases over time. Specifically, the proportion of firms with negative net income (loss firms) is below 20 percent through 1981, does not exceed 30 percent until 1985, and exceeds 40 percent for the first time in 2001. Since 2001, the proportion of loss firms exceeds 40 percent in four years and is 37.2 percent in 2015. Denis and McKeon (2016) investigate the increase in the fraction of firms with losses and document that losses are persistent, typically lasting four consecutive years. They argue that the increase in cash holdings noted in the previous section is partly due to firms raising cash to fund losses.[7]

[7] Other measures of profitability like return on assets (ROA), which includes the effect of depreciation and other noncash charges, show a similar pattern. For example, return on assets in our sample falls from 4.3 percent in 1975 to –3.3 percent in 1995 and –8.3 percent in 2015. Average and median return on assets for US corporations also decrease over our sample period, although much less so for large firms and/or in asset-weighted samples.

A substantial proportion of the decline in average operating cash flow is related to the rise of research and development spending. Recall that R&D is expensed, while capital expenditures are not. Consequently, if a firm switches from spending a fixed amount on capital expenditures to the same amount on R&D, its accounting performance worsens. To assess the importance of this effect on trends in profitability as measured by cash flow, we examine what happens when we treat R&D investment like capital expenditures: that is, we add back R&D expense to operating cash flow, so that it is also treated as capitalized. We call this measure "adjusted operating cash flow." The decline in adjusted operating cash flow over our time period is lower: from 1975 to 1995, adjusted operating cash flow averages 7.6 percent; from 1995 through 2015, it averages 6.3 percent. The equally weighted average of adjusted cash flow is never negative. However, the cash flow adjustment for R&D expenditures has less of an effect for the asset-weighted average because large firms have less R&D expenditures relative to assets than small firms.

The period from 1996 to 2015 includes the 2007–2009 Great Recession. Perhaps surprisingly, the low equally weighted averages of cash flow in the second part of our sample period are not due to the crisis years. Specifically, there are five years since 2000 when adjusted cash flow is lower than in 2008 or 2009 (and seven years when it is lower for unadjusted cash flow). Median operating cash flow to assets is higher than mean operating cash flow to assets, and is never negative; adjusted medians are the same because the median level of research and development is zero. Overall, the decrease in average cash flow is partly explained by some firms with large losses, as the drop in profitability for the typical firm is much smaller than the drop in the average.

Though performance has worsened for the average firm, the winners have done very well. One way to see this is that four new firms entered the list of the top five firms by market capitalization in 2015, relative to 1995. Specifically, Apple, Google, Microsoft, and Amazon replace AT&T, Coca Cola, General Electric, and Merck. In 2015, these four firms combined have earnings of $82.3 billion, representing 10 percent of the earnings of all public firms combined.[8] Perhaps not surprisingly, over the last 40 years, there has been a dramatic increase in the concentration of the profits and assets of US firms. Table 2 shows that in 1975, 50 percent of the total earnings of public firms is earned by the 109 top-earning firms; by 2015, the top 30 firms earn 50 percent of the total earnings of the US public firms. Even more striking, in results not separately tabulated here, we find that the earnings of the top 200 firms by earnings exceed the earnings of all listed firms combined in 2015, which means that the combined earnings of the firms not in the top 200 are negative. In 1975, the 94 largest firms own half of the assets of US public firms, but 35 do so in 2015. Finally, 24 firms account for half of the cash holdings of public firms in 1975, but 11 firms do in 2015. Table 2 also shows that the percentage of earnings accounted for by the top 100 firms almost doubles, from 48.5 percent in 1975 to 84.2 percent in 2015. For assets, cash, operating cash flow, and earnings, the share

[8]We define earnings as net income before extraordinary items, which corresponds to variable ib in Compustat.

Table 2
Concentration Statistics

	Number of firms accounting for 50% of variable in:			*Top 100 firms account for what percent of variable in:*		
Variable	*1975*	*1995*	*2015*	*1975*	*1995*	*2015*
Earnings	109	89	30	48.5%	52.8%	84.2%
Assets	94	69	35	51.1%	56.5%	66.2%
Cash	24	20	11	71.8%	73.5%	78.6%
Cash Flow	86	89	57	52.6%	52.4%	63.1%
Dividends	74	61	44	55.1%	60.6%	68.7%
Total Payouts	79	57	60	54.0%	61.4%	62.3%

Note: Detailed definitions of the variables are provided in the online Appendix at the journal website.

of the total accounted for by the top 100 firms is now at least 10 percent higher than in 1975.

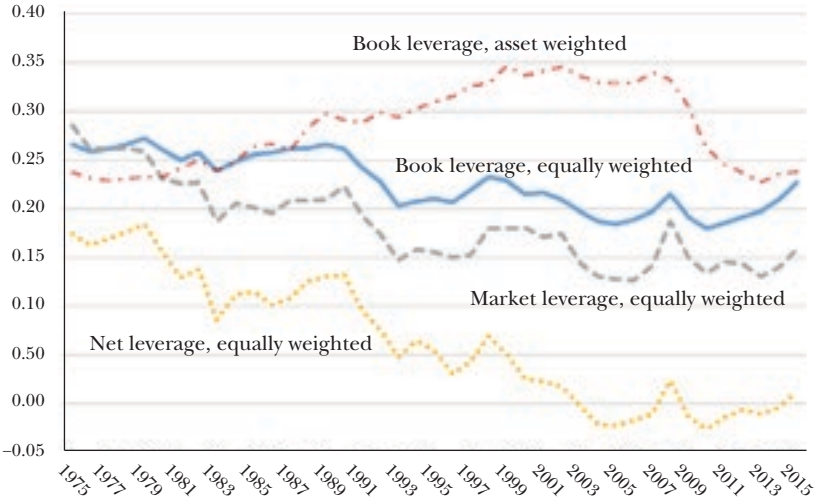
How Capital is Provided and Rewarded

As discussed earlier, US firms spend more on research and development and have less fixed assets today than they did 40 years ago. Fixed assets provide collateral against which firms can borrow, but research and development is difficult to finance with debt, as R&D in process cannot be seized by creditors if a firm gets in trouble and its value is hard to ascertain. Consequently, an increase in R&D should lead to a decrease in firm leverage. Leverage measures the importance of debt as a source of financing. The more highly levered a firm, the greater the risk of financial distress and bankruptcy, all else equal. An examination of multiple measures of leverage in Table 1 shows that leverage is lower in 2015 than in 1975. However, we saw earlier that R&D investment is more important for the equally weighted than the asset-weighted average. Therefore, the impact of increased R&D investment on leverage is expected to be more important for equally weighted measures of leverage. Our evidence supports this, in that leverage falls dramatically for an equally weighted measure of leverage that takes into account the cash holdings of firms.

Figure 4 illustrates several widely used measures of a firm's leverage. The solid line shows the equally weighted average book leverage of public corporations is slightly higher in 2015 than in 1995, but both are lower than in 1975. The asset-weighted book leverage ratio, shown by the dot-dash line, gives greater weight to large firms and tells a different story. This ratio rises substantially between 1985 and 1995, then remains high through about 2007, before dropping sharply after the financial crisis.[9]

[9] Alternative measures of leverage use the market value instead of the book value of equity. For example, the market value of assets can be calculated as total assets minus the book value of equity, plus the market

Figure 4
Equally Weighted and Asset-Weighted Book Leverage, Market Leverage, and Net Leverage as a Fraction of Total Assets from 1975 to 2015



Source: The sample is composed of the intersection of listed firms on CRSP for which Compustat data are available. Accounting data are from Compustat.
Note: The numerator of the leverage measures is long-term debt plus debt in current liabilities for equally weighted and asset-weighted book leverage and for market leverage. For net leverage, cash holdings are subtracted from the numerator. The denominator is book assets for book leverage and net leverage; for market leverage, it is book assets minus book equity plus market value of equity. Detailed variable definitions are in the online Appendix at the journal website.

Another measure of leverage examines the ratio of debt minus cash over total assets. This measure is called the "net leverage ratio," because the firm could use its cash holdings to repay its debt, and debt that is covered by cash holdings is less risky than other debt. In some ways, this net leverage ratio is a better measure of financial health than the other leverage ratios we examine. The equally weighted net leverage ratio is 0.174 in 1975. After 2003, it falls steadily and is positive in only two years, 2008 and 2015. In other words, in almost all years since 2003, the average public firm has more cash than debt. In fact, the percentage of firms with negative net leverage is 23.7 percent in 1975 and 43.1 percent in 2015. This percentage peaks at 49 percent in 2010.

value of equity. Market leverage is then the ratio of debt to the market value of assets. The decrease in leverage from 1975 to 2015 is more pronounced for the equally weighted average of market leverage than for the equally weighted average of book leverage. Regardless of whether we use the equally weighted average or the asset-weighted average, the market leverage of public firms is lower in 2015 than in 1975, and equal to or lower than what it was in 1995.

Asset-weighted net leverage (not tabulated separately here) follows a different path. It is 12.2 percent in 1975, increases to 24.7 percent in 2001, and then falls to its lowest level of 9.4 percent in 2013, ending at 11.2 percent in 2015. The asset-weighted averages of leverage and net leverage in 2015 are approximately equal to those in 1975. In other words, for large firms, leverage is not lower than in 1975, but it is lower than in all years from 1980 to 2012.

None of our leverage measures are elevated at the end of the sample period in 2015, suggesting that concerns about corporate leverage are less relevant for public firms now than at other times during the sample period. Leverage is even less of an issue now because interest rates are extremely low since the credit crisis. Hence, interest paid as a percentage of assets has never been as low during the sample period as in recent years, as shown in Table 1.

Another way to look at leverage is to examine the percent of firms that have no debt, again summarized in Table 1. The percentage of listed firms without debt increases fairly steadily from 1975, when it is 6.1 percent, to 2011, when it peaks at 18.9 percent. In 2015, it is 17.3 percent. Debt can be in the form of either publicly traded debt such as bonds, or private debt such as bank debt, but publicly available accounting data do not identify these separately. However, bank loans have become less important, according to data from the Financial Accounts of the United States published by the Federal Reserve. The Financial Accounts provide the totals of loans from depository institutions and of corporate bonds for the nonfinancial corporate sector, which includes both private and public firms. In 1975, bank loans are 56 percent of the value of corporate bonds, drop to 42 percent by 1995, and to 20 percent in 2015.[10]

In addition to debt, firms issue equity to finance themselves. Equity issuance increases the total number of shares outstanding, while repurchases decrease the total number of shares, and "net equity issuance" looks at the difference between repurchases and equity issuance. In general, smaller firms issue equity and larger firms repurchase more shares than they issue. The equally weighted average of net equity issuance divided by lagged assets follows an inverted U-shape during the last 40 years. Net equity issuance is less than 10 percent in each year in the 1970s. It increases to peak at 36.3 percent in 1996. After 1996, net equity issuance divided by assets falls. In the 2000s, it never rises above 20 percent and is lower than 10 percent in seven years. In 2015, it is 15.4 percent. An asset-weighted average gives more weight to larger firms that tend to repurchase more heavily. Asset-weighted net equity issuance is typically small but positive in the years before 2000, peaking at 2.9 percent in 2000. Since 2000, it is negative in all years but three. In 2015, it is –0.8 percent. In the 2000s, large firms are more likely to return equity to shareholders rather than raise equity from investors.

[10] These percentages are obtained by dividing item 29 (Depository institution loans n.e.c.) by item 27 (Corporate bonds) of the accounts for Nonfinancial Corporate Business of the National Accounts.

Ownership

Over the last 40 years, ownership of US publicly listed firms has changed dramatically. Corporate debt is mostly held by institutions throughout our sample period (Biais and Green 2007). However, institutional ownership of common stock is much higher now than in 1980, which is the first year in which we have data from the 13F filings at the Securities and Exchange Commission (from the Information Required of Institutional Investment Managers Form). Table 1 shows that for the first year that data is available (in 1980), 17.7 percent of outstanding shares are held by institutions, based on an equally weighted average. This percentage increases steadily and peaks at 55 percent in 2007. In 2015, this percentage is 50.4 percent. Institutions tend to prefer large firms, so institutional ownership is higher for the asset-weighted average than for the equally weighted average.

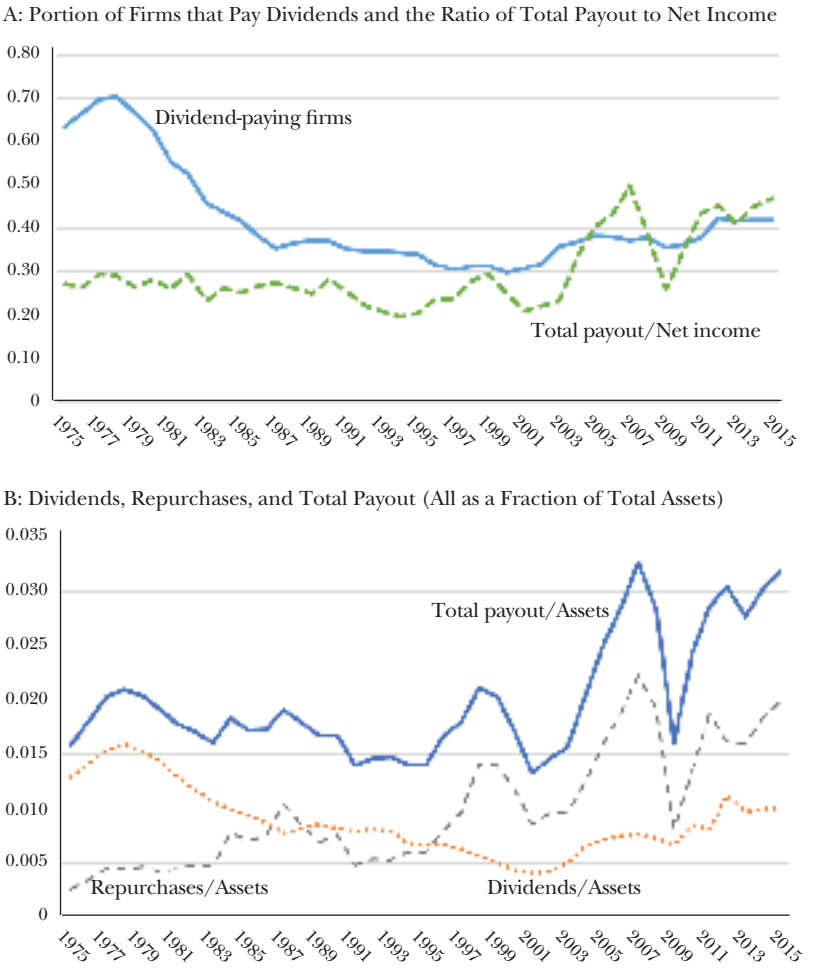
Another way that institutional ownership changes over the last 40 years is that it is now much more common for a firm to have an institutional investor who controls 10 percent or more of the shares. The percentage of US firms with a 10 percent institutional shareholder is 11.9 percent in 1980 (the first year for which data is available). This percentage increases through time, and by 1995 it is 19.5 percent. Since 2008, this percentage is typically higher than 30 percent; in 2015, 32.0 percent of firms have at least one institutional blockholder who owns 10 percent or more of the shares.

Payout Policies for Shareholders of Public Firms

Shareholders invest in equity to earn a return, which consists of current payouts and/or price appreciation. Profitable firms can use their cash flows to pay dividends, buy back shares, increase their cash holdings, or invest. Jensen's (1989) forecast of the demise of the public corporation was partly motivated by the belief that managers of public firms often retain earnings even when they cannot reinvest them profitably, which destroys shareholder wealth. Jensen (1986) called this issue the agency problem of free cash flow. He argued that public firms would tend to have payout rates that would be too low—that is, limited distributions of cash to shareholders either in the form of dividends or repurchases. In contrast, he argued that private firms can control this problem more efficiently. Yet the payout rate, defined as dividends plus repurchases as a fraction of net income, is at an all-time high in 2015. Such a high payout rate is inconsistent with worsening of agency problems, but it is consistent with a perceived lack of investment opportunities or with reduced incentives of firms to invest.

Figure 5 shows the evolution of payout rates. The percentage of dividend-paying firms follows a U-shape over the last 40 years (for discussion, see Floyd, Li, and Skinner 2015). In 1975, 63.5 percent of public firms pay dividends (as shown in Table 1). By 1995, this share falls to 34.0 percent, and it sinks to a minimum of 29.8 percent in 2000. The proportion of public firms paying dividends then

Figure 5
Dividends and Repurchases



Source: The sample is composed of listed firms on CRSP (Center for Research in Security Prices) for which Compustat data are available. Accounting data are from Compustat.
Note: Detailed variable definitions are in the online Appendix at the journal website

rebounds to 42.2 percent of listed firms in 2012, and is 41.9 percent in 2015. In 2015, the fraction of firms paying dividends is roughly one-third lower than in 1975 and one-third higher than in 2000.

Figure 5 illustrates several measures of shareholder payouts relative to the assets of firms. In 1975, the equally weighted average of dividend payments as a percentage of assets is 1.3 percent. This percentage falls to a minimum of 0.4 percent in 2000, but then rises back to roughly 1 percent in recent years. An asset-weighted

average follows the same U-shape pattern, but is slightly higher as large firms tend to pay more dividends than small firms. Total payouts to shareholders also include share repurchases. Over the past 40 years, share repurchases increase considerably (DeAngelo, DeAngelo, and Skinner 2008). In 1975, repurchases are only 0.3 percent of assets. In 1984, the Securities and Exchange Commission relaxed rules limiting repurchases by firms, and although repurchases fluctuate from year to year, they increase over time, first slowly and then more decisively. As Figure 5 shows, the equally weighted average of dividends to assets exceeds repurchases until the mid-1990s, but the relation then reverses. In asset-weighted terms, the ratio of dividends to assets is higher than the ratio of repurchases to assets until 1996. Since 1997, repurchases are higher than dividends, except in 2002 and 2003. Thus, stock repurchases are at record levels in the 2000s and extremely high in recent years. Adding together payouts from dividends and stock repurchases, the total payouts relative to assets are at historical highs in recent years, too.

Payouts can also be examined relative to the net income of the firm, rather than to the assets of the firm. The equally weighted average of total payouts in the form of both dividends and repurchases as a percentage of net income is 27.1 percent in 1975; although it sags to 20.5 percent in 1995, it is typically between 20 and 30 percent of net income from 1975 until the early 2000s. However, the payout rate then spikes to 49.9 percent of net income in 2007, decreases during the Great Recession, and then rebounds to 47.0 percent in 2015; in recent years it is higher than at any time since 1975. This evolution also occurs in the asset-weighted average. With this average, firms pay out 76.2 percent of net income in 2015, which is the fourth-highest percentage since 1975, with the three higher percentages in 2006, 2007, and 2012. By either measure, public corporations have been paying out a higher share of net income to shareholders in recent years.

Big firms account for a larger percentage of dividend payouts and a larger percentage of total payouts in 2015 than in 1975. For example, as shown in Table 2, the top 100 dividend-paying firms account for 55.1 percent of total dividends in 1975 (for additional data on the evolution of these flows, see DeAngelo, DeAngelo, and Skinner 2004). By 2015, the top 100 firms account for 68.7 percent of total dividends. The same increase in concentration has taken place for total payouts, but the increase is more muted as the top 100 firms account for 54 percent in 1975 and 62.3 percent in 2015.

How to Make Sense of Our Results

The changes we document will be topics of research for years to come, but at this stage, in the absence of consensus on the explanations, it is useful to consider some leading possibilities that have either been advanced already or are worth considering.

Let's begin with a benign potential explanation. In a market economy, resources are constantly reallocated from less-efficient firms to more-efficient firms. Hence,

at times, this reallocation will naturally lead to consolidation, with less-efficient firms being acquired by more-efficient firms. This process is reinforced if larger firms have an efficiency advantage because of their size. In this case, it will not be surprising to see the number of firms fall and the larger firms survive. In this view, the US economy entered a period of consolidation in the mid-1990s and, hence, we have larger but fewer public firms.

One reason to be skeptical of this benign explanation is that the consolidation is concentrated within the universe of public firms. If consolidation has nothing to do with being a public firm, we should see the total number of firms decreasing, whether firms are public or private. We don't. The United States has become an economy dominated by service industries, and so a good way to demonstrate this is to look at the service industries. Even though the number of firms in the service industries increases by 30 percent from 1995 to 2014 and employment increases by 240 percent, the number of public firms falls by 38 percent. A similar evolution occurs in the finance industry, in which the number of firms increases by 18.7 percent from 1995 to 2014, but over the same time the number of listed firms falls by 42.3 percent. Further, Doidge, Karolyi, and Stulz (2017) show that the propensity of firms to be listed—which they define as the percentage of public firms in the population of all firms—falls across all firm-size categories when size is measured by employment. The efficient consolidation view is also challenged by evidence suggesting that mergers in recent years do not have efficiency gains, but instead the gains have come from larger markups (Blonigen and Pierce 2016). Grullon, Larkin, and Michaely (2016) argue that this consolidation seems to be partly the result of a relaxation of antitrust enforcement, and so it is occurring because of mergers that might not have taken place earlier on antitrust grounds.

The drop in the propensity to be listed suggests that there is a problem with being a public firm. Many have argued that the regulatory burden associated with being public increased as a result of the Sarbanes–Oxley Act of 2002 and that, as a result, fewer firms want to be public and many of them have exited public markets. The problem with this explanation is two-fold. First, the drop in the number of public firms predates the regulatory changes of the early 2000s, so these changes can only be a partial explanation. Second, as discussed earlier, the fraction of firms going private is small compared to the fraction of firms that are no longer listed because of mergers. However, the topic of Sarbanes–Oxley does highlight a problem with public firms. In the United States, corporate law is governed by state of incorporation, but public firms are subject to federal securities laws. As a result, Congress can regulate public firms in ways that it cannot regulate private firms. For instance, concerns about conflict minerals led to Section 1502 of the Dodd–Frank Wall Street Reform and Consumer Protection Act, which mandates disclosure by public firms of whether their supply chain uses such minerals. Such a requirement has an asymmetric effect, because private firms do not face the same requirement.

Our data show that the fraction of small public firms has dropped dramatically. Gao, Ritter, and Zhu (2013) document that the drop in initial public offerings is particularly acute among small firms. Why are public markets no longer welcoming

for small firms? We already saw that research and development investments have become more important. Generally, R&D is financed with some form of equity rather than debt, at least in early stages before a firm has accumulated lucrative patents. Raising equity in public markets to fund R&D can be difficult. Investors want to know what they invest in, but the more a firm discloses, the more it becomes at risk of providing ammunition to its competitors. As a result, R&D-intensive firms may be better off raising equity privately from investors who then have large stakes. These firms can explain their R&D program in greater detail to such investors without worrying as much about providing information to the competition.

There are several additional potential explanations for why small firms are staying out of public markets: changes in financial markets and intermediation, increased economies of scope, increased concentration, and changes in how firm activities are organized. The financial markets and intermediation explanation has two parts. First, public markets have become dominated by institutional investors. As a result, financial institutions and exchanges cater more to the demands of these investors. Investing in really small firms is unattractive for institutional investors, because they cannot easily invest in a small firm on a scale that works for them. As a result, small firms receive less attention and less support from financial institutions. This makes being public less valuable for these firms. Second, developments in financial intermediation and regulatory changes have made it easier to raise funds as a private firm. Private equity and venture capital firms have grown to provide funding and other services to private firms. The internet has reduced search costs for firms searching for investors. As a result, private firms have come to have relatively easier access to funding.

Gao, Ritter, and Zhu (2013) advance the economies of scope hypothesis. According to this hypothesis, small firms have become less profitable and less able to grow on a stand-alone basis, but are more profitable as part of a larger organization that enables them to scale up quickly and efficiently. Thus, small firms are better off selling themselves to a large organization that can bring a product to market faster and realize economies of scope. This dynamic arises partly because it has become important to get big quickly as technological innovation has accelerated. Globalization also means that firms must be able to access global markets quickly. Further, network and platform effects can make it more advantageous for small firms to take advantage of these effects by being acquired. This hypothesis is consistent with our evidence that the fraction of exchange-listed firms with losses has increased and that average cash flows for smaller firms have dropped. Gao, Ritter, and Zhu (2013) and others also show that many mergers do involve small firms, so small firms do indeed choose to be acquired rather than grow as public firms.

The increased concentration we document could also make it harder for small firms to succeed on their own, as large established firms are more entrenched and more dominant. It could be that private firms can grow more easily before they attempt to reach a national market but face more daunting obstacles if they try to become public and compete with the larger, more established firms. Further, it may be harder for smaller firms to compete and stay independent in a world where intellectual property has become so important, as these firms may find it difficult

to acquire the rights to patents that allow them to grow and exploit their own intellectual property. Hence, the growing importance of research and development may itself lead to a world where competition is more limited.

Davis (2016) argues that it has become easier to put a new product on the market without hard assets. Entrepreneurs can rent and can outsource. For instance, Vizio rapidly overtook Sony in terms of television sales with less than 200 employees and not producing anything in house. Netflix rents server farms from Amazon. When all the pieces necessary to produce a product can be outsourced and rented, a firm can bring a product to market without large capital requirements. Hence, the firm does not need to go public to raise vast amounts of equity to acquire the fixed assets necessary for production. The top five firms in 2015 have relatively few employees. Ford's largest production facility in the 1940s, the River Rouge complex, employed more than 100,000 workers at its peak. Of today's largest US firms, only Amazon has substantially more employees than that complex at its peak. With this evolution, there is no point in going public, except to enable owners to cash out.

These explanations imply that there are fewer public firms both because it has become harder to succeed as a public firm and also because the benefits of being public have fallen. As a result, firms are acquired rather than growing organically. This process results in fewer thriving small public firms that challenge larger firms and eventually succeed in becoming large. A possible downside of this evolution is that larger firms may be able to worry less about competition, can become more set in their ways, and do not have to innovate and invest as much as they would with more youthful competition. Further, small firms are not as ambitious and often choose the path of being acquired rather than succeeding in public markets. With these possible explanations, the developments we document can be costly, leading to less investment, less growth, and less dynamism.

Conclusion

US public firms are very different now compared to 1975 or 1995: fewer, larger, older, less-profitable, with more intangible capital, less investment, and other changes. The US firms that remain public are mostly survivors. Few firms want to join their club. A small number of firms account for most of the market capitalization, most of the earnings, most of the cash, and most of the payouts of public firms. At the industry level, revenues are more concentrated, so fewer public firms are competing for customers. A large fraction of firms do not earn profits every year and that fraction is especially large in recent years, which helps to explain the high level of delists. Accounting standards do not reflect the importance of intangible assets for listed firms, which may make it harder for executives to invest for the long run.

The key argument of Jensen (1989) in his forecast of the demise of the public firm is that the public firm is beset by agency problems. The fact that US firms pay out more to shareholders now than at any time since 1975 seems inconsistent with the view that the central agency problem involves managers retaining resources

internally instead of paying them out to shareholders. However, Jensen's prediction of the rise of private equity has proven to be on the mark. The rise of private equity may be one of the contributing factors for why so few firms choose to participate in the public markets.

Since the 1997 peak in the number of listed public firms, the number of firms has dropped sharply while revenues have become more concentrated. Even though Tobin's *q* is high, firms invest less, and they have record payouts. Public firms as a whole are repurchasing more equity than they issue in most years since 2000. It appears in that firms are less dependent on public markets to raise capital to finance investments. It may be in the best interests of shareholders for firms to behave that way, but the end result is likely to leave us with fewer public firms, who gradually become older, slower, and less ambitious. Consequently, fewer new private firms are born, as the rewards for entrepreneurship are not as large. And those firms that are born are more likely to lack ambition, as they aim to be acquired rather than to conquer the world.

■ *We are grateful for discussions with Harry DeAngelo and comments from the editors, Andrei Gonçalves, Andrew Karolyi, Steve Kaplan, and Jay Ritter.*

References

Bates, Thomas W., Kathleen M. Kahle, and Rene M. Stulz. 2009. "Why Do U.S. Firms Hold So Much More Cash than They Used To?" *Journal of Finance* 64(5): 1985–2021.

Biais, Bruno, and Richard C. Green. 2007. "The Microstructure of the Bond Market in the 20th Century." http://repository.cmu.edu/tepper/134/.

Blonigen, Bruce A., and Justin R. Pierce. 2016. "Evidence for the Effects of Mergers on Market Power and Efficiency." NBER Working Paper 22750.

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2009. "Intangible Capital and U.S. Economic Growth." *Review of Income and Wealth* 55(3): 661–85.

da Silveira, Giovani J. C. 2014. "An Empirical Analysis of Manufacturing Competitive Factors and Offshoring." *International Journal of Production Economics* 150: 163–73.

Davis, Gerald F. 2016. *The Vanishing American Corporation: Navigating the Hazards of a New Economy*. Berrett-Koehler Publishers.

DeAngelo, Harry, Linda DeAngelo, and Douglas J. Skinner. 2004. "Are Dividends Disappearing? Dividend Concentration and the Consolidation of Earnings." *Journal of Financial Economics* 72(3): 425–56.

DeAngelo, Harry, Linda DeAngelo, and Douglas J. Skinner. 2008. "Corporate Payout Policy." *Foundations and Trends in Finance* 3(2–3): 95–287.

Denis, David J., and Stephen B. McKeon. 2016. "Persistent Operating Losses and Corporate Financial Policies." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2881584.

Doidge, Craig, G. Andrew Karolyi, and René M. Stulz. 2013. "The U.S. Left Behind? Financial Globalization and the Rise of IPOs Outside the U.S." *Journal of Financial Economics* 110(3): 546–73.

Doidge, Craig, G. Andrew Karolyi, and René M. Stulz. 2017. "The U.S. Listing Gap." *Journal of Financial Economics* 123(3): 464–87.

Eisfeldt, Andrea L., and Dimitris Papanikolaou. 2014. "The Value and Ownership of Intangible Capital." *American Economic Review* 104(5): 189–94.

Falato, Antonio, Dalida Kadyrzhanova, and Jae W. Sim. 2013. "Rising Intangible Capital, Shrinking Debt Capacity, and the U.S. Corporate Savings Glut." Finance and Economics Discussion Paper 2013–67, Board of Governors of the Federal Reserve System.

Floyd, Eric, Nan Li, and Douglas J. Skinner. 2015. "Payout Policy through the Financial Crisis: The Growth of Repurchases and the Resilience of Dividends." *Journal of Financial Economics* 118(2): 299–316.

Gao, Xiaohui, Jay R. Ritter, and Zhongyan Zhu. 2013. "Where Have All the IPOs Gone?" *Journal of Financial and Quantitative Analysis* 48(6): 1663–92.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely. 2016. "Are U.S. Industries Becoming More Concentrated?" Available at SSRN: https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2612047.

Hathaway, Ian, and Robert Litan. 2014. *The Other Aging of America: The Increasing Dominance of Older Firms*. Washington, DC: Brookings Institute.

Jensen, Michael C. 1986. "Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers." *American Economic Review* 76(2): 323–29.

Jensen, Michael C. 1989. "Eclipse of the Public Corporation." *Harvard Business Review* 67 (September–October): 61–74.

Levine, Ross. 1997. "Financial Development and Economic Growth: Views and Agenda." *Journal of Economic Literature* 35(2): 688–726.

Loderer, Claudio, René M. Stulz, and Urs Waelchi. Forthcoming. "Firm Rigidities and the Decline of Growth Opportunities." *Management Science*.

Weild, David, and Edward Kim. 2010. *Market Structure is Causing the IPO Crisis—And More.* London: Grant Thornton International.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 89–112

The Agency Problems of Institutional Investors

Lucian A. Bebchuk, Alma Cohen, and Scott Hirst

Financial economics and corporate governance have long focused on the agency problems between corporate managers and shareholders that result from the dispersion of ownership in large publicly traded corporations. In this paper, we focus on how the rise of institutional investors over the past several decades has transformed the corporate landscape and, in turn, the governance problems of the modern corporation. The rise of institutional investors has led to increased concentration of equity ownership, with most public corporations now having a substantial proportion of their shares held by a small number of institutional investors. At the same time, these institutions are controlled by investment managers, which have their own agency problems vis-à-vis their own beneficial investors. These agency problems are the focus of our analysis.

■ *Lucian A. Bebchuk is the James Barr Ames Professor of Law, Economics, and Finance and the Director of the Program on Corporate Governance, both at Harvard Law School, Cambridge, Massachusetts. Alma Cohen is Professor of Empirical Practice at Harvard Law School and Associate Professor, Eitan Berglas School of Economics, Tel-Aviv University, Tel-Aviv, Israel. Scott Hirst is Research Director of the Program on Institutional Investors and Lecturer on Law, both at Harvard Law School, Cambridge, Massachusetts. Bebchuk is a Research Associate and Cohen is a Faculty Fellow, National Bureau of Economic Research, Cambridge, Massachusetts. Their emails are bebchuk@law.harvard.edu, alcohen@law.harvard.edu, and shirst@law.harvard.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.89

doi=10.1257/jep.31.3.89

We develop an analytical framework for understanding the agency problems of institutional investors. We apply this framework to examine the agency problems and behavior of several key types of investment managers, including those that manage mutual funds—both index funds and actively managed funds—and activist hedge funds.

We identify several drivers of agency problems that afflict the decisions of investment managers of either passive index funds, active mutual funds, or both. First, such investment managers generally capture only a small fraction of the benefits that results from their stewardship activities while bearing the full cost of such activities. Further, competition with other investment managers is typically insufficient to eliminate these agency problems. Finally, investment managers may be further influenced by private incentives, such as their interest in obtaining business from corporations, that encourage them to side excessively with managers of corporations.

We show that index funds have especially poor incentives to engage in stewardship activities that could improve governance and increase value. Accordingly, while the rise of index funds benefits investors and the economy by reducing the costs of financial intermediation, this trend also has systemwide adverse consequences on governance.

Activist hedge funds have substantially better incentives than managers of index funds or active mutual funds. While their activities may partially compensate, we show that they do not provide a complete solution for the agency problems of other institutional investors.

We recognize that well-meaning investment managers of index funds and active mutual funds may sometimes make stewardship decisions that are superior to those suggested purely by their incentive calculus. Our focus, however, is on understanding the structural incentive problems that should be recognized in assessing the current governance landscape.

There is a growing recognition by researchers, capital market participants, and public officials that investment fund managers are imperfect agents for those investing in their funds, and there is now significant literature on this problem. Our analytical framework contributes by identifying the direction and manner in which the behavior of investment fund managers can be expected to deviate from the interests of their beneficial investors. For example, by demonstrating that the agency problems of institutional investors can be expected to lead them to underinvest in stewardship and side excessively with corporate managers, we show that concerns about the existence of such agency problems provide little basis for weakening shareholder rights or impeding shareholder action.

Furthermore, our analysis also generates insights on a wide range of policy questions and provides a framework for future work. We conclude by offering implications in a number of areas: disclosure by institutional investors and regulation of their fees; stewardship codes; the rise of index investing; proxy advisors; hedge fund and wolf pack activism; the allocation of power between corporate managers and shareholders; and others.

The Rise of Institutional Investors

In their classic work on the separation of ownership and control, Berle and Means (1932) introduced the problem of publicly traded companies with widely dispersed ownership. In such situations, Berle and Means explained that, "[a]s his personal vote will count for little or nothing at the meeting … the stockholder is practically reduced to the alternative of not voting at all or else of handing over his vote" to the proxy committee, appointed by existing management, who can "virtually dictate their own successors" (p. 87). Because dispersed shareholders can thus be expected to be rationally apathetic, managers will be relatively unconstrained in their actions, which Berle and Means refer to as "management control" of the corporation.

Furthermore, Berle and Means (1932) documented that a significant proportion of publicly traded corporations have a sufficiently broad dispersion of shareholders to be classified as management-controlled. For example, Berle and Means (pp. 107–109, table XII, panel G) show that, of the largest 200 corporations in 1930 that they listed as being controlled by hired managers (rather than run directly by owners), the aggregate percentage of the corporation's equity owned by the corporation's largest 20 shareholders had a mean of 10.55 percent (median of 10.6 percent).

Some classic articles by financial economists, following Berle and Means (1932), assume that shareholders of publicly traded firms are "atomistic" and have no incentive to seek governance improvements in the firms in which they own shares (for example, Grossman and Hart 1980; Shleifer and Vishny 1986). Given the practical infeasibility of such shareholder activities in the Berle–Means corporation, some researchers have focused on how other mechanisms, such as the discipline of the market for corporate control (Manne 1965), stock ownership by managers (Demsetz 1983), or price pressure due to sale of shares by investors seeking to exit underperforming companies (Admati and Pfleiderer 2009) could constrain the agency problems of managers and thereby make up for the lack of direct shareholder effort to improve governance.

Berle and Means (1932, p. 47) argued that "[d]ispersion in the ownership of separate enterprises appears to be inherent in the corporate system. It has already proceeded far, it is rapidly increasing, and appears to be an inevitable development." However, the trend toward dispersion has been reversed in subsequent decades by the rise of institutional investors. The rise of institutional investors has been driven by investor recognition of the value of low-cost diversification and encouraged by favorable regulatory and tax treatment. Whereas institutional investors held 6.1 percent of outstanding corporate equity in 1950 (Tonello and Rabimov 2010), they held 63 percent of outstanding public corporate equity in 2016 (Board of Governors of the Federal Reserve System 2016, p. 130). Furthermore, because institutional investors aggregate the assets of a vast number of individuals, each institutional investor can hold large positions in many publicly traded companies.

Table 1
Institutional Ownership of the 20 Largest US Corporations

		Percentage owned by largest holders		
	Corporation	*Largest 5*	*Largest 20*	*Largest 50*
1.	Apple Inc.	17.5%	26.8%	35.4%
2.	Microsoft Corp.	20.5%	33.1%	43.2%
3.	Exxon Mobil Corp.	17.8%	27.1%	35.2%
4.	Johnson & Johnson	19.0%	30.3%	40.5%
5.	General Electric Co.	17.5%	28.0%	37.3%
6.	AT&T Inc.	19.0%	28.8%	37.4%
7.	Wells Fargo & Co.	24.9%	40.2%	51.0%
8.	Verizon Communications Inc.	20.1%	32.9%	43.7%
9.	Procter & Gamble Co.	18.4%	28.3%	38.2%
10.	JPMorgan Chase & Co.	19.5%	34.7%	47.1%
11.	Pfizer Inc.	18.7%	32.1%	45.1%
12.	Chevron Corp.	21.6%	33.9%	43.6%
13.	Coca-Cola Co.	26.6%	39.9%	48.6%
14.	Visa Inc.	23.8%	41.7%	56.3%
15.	Home Depot Inc.	24.4%	37.4%	49.1%
16.	Disney (Walt) Co.	17.9%	29.6%	39.1%
17.	Merck & Co.	26.1%	38.4%	50.1%
18.	Philip Morris International	24.8%	40.9%	52.1%
19.	Intel Corp.	20.2%	32.9%	44.6%
20.	Cisco Systems Inc.	18.8%	32.2%	45.7%
Mean		**20.8%**	**33.4%**	**44.2%**
Median		**19.8%**	**32.9%**	**44.2%**

Source: FactSet Ownership database (by FactSet Research Systems).
Note: The table shows the aggregate ownership of the largest holders of the largest 20 US corporations by market capitalization as of June 30, 2016, excluding controlled corporations.

As a result of the rise of institutional investors, the scenario of dispersed ownership described by Berle and Means (1932) no longer approximates reality, not even for the largest publicly traded corporations. Table 1 lists the largest 20 US corporations by market capitalization as of June 30, 2016 (excluding controlled corporations), and the aggregate percentage of the stock of each corporation owned by their largest 5, 20, and 50 institutional investors.[1]

As Table 1 shows, current share ownership is significantly more concentrated than the level described by Berle and Means (1932). Indeed, because the figures in Table 1 exclude large holdings by noninstitutional investors, they likely underestimate the degree to which shares are concentrated among investors with significant holdings. Even among the largest 20 corporations, the largest 20 institutional

[1] Investment advisers that manage multiple mutual funds generally have corporate governance staff that cast votes in the same way for each fund and undertake stewardship on behalf of each fund. Accordingly, for the purposes of these calculations, we group the shareholdings of the mutual funds managed by each investment manager as a single "institutional investor."

investors in 2016 had mean ownership of 33.4 percent (and similar median ownership of 32.9 percent), more than three times the figure reported by Berle and Means (1932); in each of the 20 corporations, the largest 20 institutional investors own more than 25 percent. Furthermore, among these very large public corporations, the percentage owned by the largest 50 institutional investors has a mean of 44.2 percent (the median is also 44.2 percent). The increase in concentration is perhaps most vivid when looking at the aggregate percentage owned by the largest five shareholders, which has a mean of 20.8 percent (median of 19.8 percent) and is above 17 percent in each of the 20 largest US corporations.

Data from ISS Voting Analytics shows that the mean percentage of shares outstanding voted at the 2015 annual meetings of these corporations for the election of directors was 68.7 percent (median of 70.8 percent). The largest 50 institutional investors thus cast a substantial majority of the votes at these annual meetings.

Thus, large institutional shareholders hold sufficiently sizable positions in each large corporation to have a non-negligible effect on the outcomes of shareholder votes. Moreover, these shareholders recognize that many of their fellow shareholders are similarly non-atomistic. Of course, because the benefits of each shareholder's actions will be shared with fellow shareholders, it will still be privately optimal for each shareholder to underspend on stewardship. However, given the current concentrated ownership of publicly traded corporations, if each shareholder were solely investing its own money, it would no longer be rational for all shareholders to be rationally apathetic. On the contrary, given that some stewardship involves limited costs and can generate significant increases in value, it is likely to be privately optimal for some shareholders with significant holdings to undertake such activities.

As a result of these changes, the prospects for stewardship by shareholders are substantially better today than in Berle–Means corporations. Institutional investors participate in corporate voting, and there is empirical evidence that the presence of institutional investors influences how corporations are governed (for example, Hartzell and Starks 2003; Aghion, Van Reenen, and Zingales 2013). Institutional investors therefore provide constraints on agency problems in their portfolio companies that dispersed shareholders in Berle–Means corporations were unable to accomplish.

However, investment managers invest other people's money. Thus, the question arises whether their stewardship decisions would be the same as those that they would make if they were solely investing their own money. Below we analyze the agency problems that could lead these investment managers to deviate from the stewardship decisions that would be optimal for their beneficial investors. These agency problems limit the extent to which our corporate governance system is able to benefit from the increased concentration of shareholdings, and are a key impediment to improving the governance of publicly traded corporations.

Stewardship by Investment Managers

Investment Funds, Active and Passive

By investment funds we refer to funds that pool together the assets of many individuals and entities and invest them in a diversified portfolio of securities. The category of investment funds includes open-end mutual funds, closed-end mutual funds, exchange-traded funds, and other similar funds. Most of these investment funds are technically "investment companies," as defined and regulated by the Investment Company Act of 1940. Given our emphasis on corporate governance, we naturally focus on funds that invest in equity securities. Investment funds are the most important category of institutional investors and represent most of the assets held by institutional investors.

Investment funds generally enter into contracts with organizations, referred to in US securities regulations as "investment advisers," to manage the porfolios of investment funds. We will refer to these organizations as "investment managers."

Investment funds focusing on equity securities can be categorized by their investing strategy into those that actively manage their portfolio and those that passively invest by matching their portfolio weightings of corporations to those of an underlying equity index. We refer to the latter, which include both open-end mutual funds and exchange-traded funds, as index funds. Most mutual fund managers operate a number of mutual funds, often referred to collectively as a "mutual fund family." While most mutual fund families include both actively managed funds and index funds, mutual fund families predominantly operate one or the other kind of investment fund.

The index fund market is dominated by three investment managers: BlackRock, Vanguard, and State Street Global Advisors (sometimes referred to as the "Big Three"). These investment managers have assets under management of $3.1 trillion, $2.5 trillion, and $1.9 trillion, respectively (Diamond 2016). The largest investment managers of actively managed funds include Fidelity Investments and the Capital Group, both of which have more than $1 trillion in assets under management.

We pay particular attention to index funds because their share of the market for managed investments has increased significantly in recent years, a trend that is expected to continue. The move towards index funds is driven by the growing recognition of their low costs and tax advantages, and the evidence that they outperform most actively managed equity mutual funds (French 2008). Passively managed funds increased from 1 percent of total fund assets in 1984 to 12.6 percent in 2006 (French 2008), and the move from active to passive funds has continued since then. From 2013 to 2016, investors added $1.3 trillion to passive mutual funds and exchange-traded funds (Tergesen and Zweig 2016).

The rise of index investing has benefits in reducing the costs of intermediation borne by investors; as of the end of 2015, the asset-weighted average net expense ratio was only 0.12 percent for US equity index funds, compared

to 0.79 percent for actively managed US equity funds (Oey and West 2016, p. 6). We recognize this benefit to investors, but wish to stress a systemic cost of index funds. As we discuss below, while agency problems afflict the stewardship activities of all investment funds, they are likely to be especially acute for index funds.

Stewardship

Our focus is on those decisions of investment managers that relate to the stewardship of companies in their portfolio. Stewardship by investment managers can take several forms. Most investment funds are required to vote at shareholder meetings on director elections and management and shareholder proposals, and to have an internal process for making voting decisions. Thus, not voting, or voting in a patently uninformed manner, is not an option for investment managers. Stewardship therefore requires monitoring of corporate managers and other information gathering in order to inform voting, engagement, and other stewardship activities. Investment managers can nominate candidates for election as directors or put forward shareholder proposals, and they can communicate with the corporation, or with other shareholders, about such matters. While stewardship may also relate to environmental and social matters that affect investors (for example, Hirst 2016), our focus in this paper is on stewardship decisions that affect beneficial investors only through their effect on the financial value of the managed portfolio.

Stewardship decisions can be split into two parts: 1) spending decisions regarding how much to expend on stewardship; and 2) qualitative decisions regarding which way to vote or which positions to take in communications with corporate managers and other shareholders.

Like all organizations with multiple employees, investment managers have their own internal agency problems. Our analysis can be thought of as analyzing the incentives that would shape the stewardship strategies that the leaders of investment managers would pursue, for example, choices regarding the resources to provide for corporate governance and proxy voting units and setting the general policy and approach of such units.

Because the voting and stewardship decisions of mutual fund families are commonly concentrated in a single corporate governance department or proxy voting department of the investment manager, the stewardship incentives of investment managers with different types of funds are a composite of the different incentives we identify below for the different types of investment funds.

Sources of Agency Problems

The Benchmark Scenario: Decisions that Maximize Portfolio Value

Let us consider a hypothetical scenario with no agency problems in managing such investments. For instance, imagine that each of the positions were those of sole owners that owned and managed 100 percent of each investment. In this case,

the decisions made would be ones that maximize the value of the owners' wealth. More specifically, suppose that some stewardship activity will cost C and will increase the value of the position by ΔV. Then, in the benchmark, no-agency scenario, the stewardship activity will be undertaken if $C < \Delta V$.[2]

For large equity positions, like those that investment managers hold in many companies, the no-agency-costs scenario would often justify meaningful investments in stewardship activities. If an investor had a $1 billion investment in a given portfolio company, and investment in certain stewardship activities would increase the value of the company by 0.1 percent, then the investor would have an incentive to spend up to $1 million on stewardship to bring about this change. We note that each large mutual fund family holds positions exceeding $1 billion in value in a large number of public companies; data from the FactSet Ownership database shows that, as of December 31, 2016, BlackRock, Capital Research, Fidelity, State Street Global Advisors, and Vanguard each held such positions at a substantial proportion of corporations in the S&P 500 index.

In many cases, stewardship decisions may be merely qualitative, and not involve additional cost. This is commonly the case when investment managers decide how to cast a vote or what position to take in interactions with corporate managers or fellow shareholders. Suppose that voting or otherwise taking a position against the outcome management prefers would change the value of the position by ΔV, where ΔV can be positive or negative. In such a case, in the no-agency-cost benchmark scenario, the investor should make a choice against managers' preferences whenever ΔV is positive.

Capturing Only a Small Fraction of the Benefit

We now turn to the decisions that the investment manager would find privately optimal. Although we will later relax these assumptions, we will initially take as given the size of fees charged by investment managers and the size of the portfolio managed.

One key source of agency problems is that investment managers bear the costs of stewardship activities, but capture only a small fraction of the benefits they create. Under existing regulations governing mutual funds, investment managers cannot charge their personnel and other management expenses directly to the portfolio. For example, if an investment manager were to employ staff fully dedicated to stewardship of a single corporation, or if an investment manager were to conduct a proxy fight in opposition to incumbent managers, it would have to cover those expenses itself, out of the fee income it receives from investors.

At the same time, the benefits from stewardship flow to the portfolio. Mutual fund managers and investment managers of other similarly structured funds are not permitted to collect incentive fees on increases in the value of their portfolio but

[2] In developing our analytical framework, we draw upon the model in Bebchuk and Neeman (2010), which explains how the decisions that institutional investors make with respect to lobbying regarding investor protection levels differ from the decisions that would be optimal for the beneficial investors in those funds.

may only charge fees that are calculated as a percentage of assets under management. Let α be the fraction of assets under management that an investment manager charges as fees. Therefore, α is also fraction of the increase in the value of a portfolio company that an investment fund will be able to capture, in present value terms, from additional fees. The value of α is likely to be small given that the asset-weighted average net expense ratio for US equity index funds was 0.12 percent as of December 31, 2015 (Oey and West 2016). It would not be in the interests of the investment manager to spend an amount C that would produce a gain of ΔV to the portfolio if C is larger than $\alpha \times \Delta V$. Thus, in this setting, agency problems would lead to underspending on stewardship, precluding efficient expenditure, whenever:

$$\alpha \times \Delta V < C < \Delta V$$

To illustrate this wedge, reconsider the example above of an investment manager of an index fund that holds a $1 billion investment in a portfolio company whose value could increase by $1 million as a result of certain stewardship activities. If the investment manager could expect additional fees with a present value of 0.12 percent from the changes in the value of the position, it would be willing to take such actions only if their cost was below $1,200, compared to $1 million in the no-agency-costs scenario.

Although investment managers of actively managed funds charge higher fees, because those fees are still a very small fraction of the investment, they will have only slightly higher incentives to spend on stewardship. If such an investment manager received additional fees of 0.79 percent of the change in the value of the position—the asset-weighted average net expense ratio for actively-managed US equity mutual funds as of December 31, 2015 (Oey and West 2016)—then it would be willing to take such actions only if their cost was below $7,900. Thus, managers of active mutual funds still have strong incentives to spend much less on stewardship than would be value-maximizing for their portfolio.

The Limits of Competition: Index Funds

Thus far, our analysis has assumed that investment managers take their fees and assets under management as given when making stewardship decisions. By relaxing this assumption, we now consider whether the desire to improve performance and attract additional funds might counter the distortions identified above and lead investment managers to make additional investments in stewardship that would be portfolio-value-maximizing.

In examining this question, it is important to recognize that what matters for attracting assets under management (and thereby increasing future fee revenue) is not the absolute performance of the investment manager, but its performance relative to alternative investment opportunities. Potential investors in equity mutual funds can be expected to judge the investment manager's performance relative to an equity index, or relative to other comparable equity mutual funds. As a result, in

many cases, the consideration of improving relative performance would not provide any incentives to improve stewardship decisions.

In particular, this is the situation in the important case of the investment managers of passively managed index mutual funds. If the investment manager of a certain mutual fund that invests according to a given index increases its spending on stewardship at a particular portfolio company and thereby increases the value of its investment in that company, it will also increase the value of the index, so its expenditure would not lead to any increase in the performance of the mutual fund relative to the index. Nor would it lead to any increase relative to the investment manager's rivals that follow the same index, as any increase in the value of the corporation would also be captured by all other mutual funds investing according to the index, even though they had not made any additional expenditure on stewardship.

Thus, if the investment manager were to take actions that increase the value of the portfolio company, and therefore also the portfolio that tracks the index, doing so would not result in a superior performance that could enable the manager to attract funds currently invested with rival investment managers. Such decisions would also not enable the investment manager to increase fees relative to rivals tracking the same index, as such rivals would offer the same gross return without the increased fees. Accordingly, for managers of index funds, a desire to improve relative performance would not provide *any* incentives that could counter tendencies that the investment manager might otherwise have to underspend on stewardship and to side with corporate managers more often than is optimal for the investment managers' beneficial investors.

It could be argued that the inability of index funds to attract additional investors by increasing stewardship spending implies that the existing equilibrium is optimal. However, our analysis indicates that this equilibrium is due to a collective action problem. The beneficial investors of an index fund would be better served by having the fund increase stewardship spending up to the level that would maximize the portfolio value, even if the fund increased its fees to fund this spending. However, if the index fund were to raise its fees and improve its stewardship, each individual investor in the fund would have an incentive to switch to rival index funds. That is, a move by any given index fund manager to improve stewardship and raise fees would unravel, because its investors would prefer to free-ride on the investment manager's efforts by switching to another investment fund that offers the same indexed portfolio but without stewardship or higher fees.

The Limits of Competition: Actively Managed Funds

Turning to actively managed funds, it is important to recognize that there is evidence that many of these funds are, to varying extents, "closet indexers" whose holdings substantially overlap with their benchmark index, deviating only by underweighting and overweighting certain stocks (Cremers and Petajisto 2009). For an actively managed fund that is to some extent a closet indexer, a desire to improve relative performance would provide no incentives to move stewardship decisions toward optimality for any of the portfolio companies where the company's

weighting in the investment fund's portfolio is approximately equal to its weighting in the index; improving the value of those portfolio companies would not enhance performance relative to the index.

Furthermore, for all the corporations that are underweight in the portfolio relative to the index, enhancing the value of the corporation would actually *worsen* the investment manager's performance relative to the index. For corporations that are underweight in the portfolio, the consideration of increasing relative performance does not provide any incentive to enhance the value of these corporations; on the contrary, this consideration weighs *against* trying to do so.

Thus, the desire to improve relative performance could only provide an actively managed fund with incentives to improve value in those corporations that are overweight in the portfolio compared to the index. Even for such corporations, the extent to which improving the value of the corporation would improve fund performance will depend on the extent to which the corporation is overweight in the portfolio.

Consider a portfolio company that constitutes 1 percent of the benchmark index and 1.2 percent of the investment fund. In this case, any increase in the value of the portfolio company will be substantially shared by rival funds that track the index at least partly. Indeed, the increase in value of the portfolio company will worsen the performance of the investment fund relative to rival funds that are more overweight with respect to the portfolio company. Thus, even for companies that are overweight within the portfolio of the investment fund relative to the index, the impact of the desire to improve relative performance would be diluted by the presence of the company in the benchmark index and in the portfolios of rival funds.

Furthermore, as discussed above, in most cases actively managed funds are part of mutual fund families composed of a number of mutual funds, and stewardship decisions are commonly made for all these investment funds by the fund family's governance or proxy voting group. In such a case, the fact that a given actively managed fund is overweight in a particular corporation might be offset by the fact that other actively managed funds within the same fund family might be underweight. The investment manager of the fund family will have an incentive to bring about an increase in value only if its actively managed funds are on the whole overweight in this corporation, and the incentive will be diluted to the extent that any gains will be shared by other mutual fund families.

In addition, an interest in improving their relative performance might also push investment managers in the opposite direction, and thereby exacerbate rather than alleviate distortions in stewardship decisions. Ke, Petroni, and Yu (2008, p. 855) describe evidence that some institutional investors value "direct access to companies' management," presumably because they believe that, notwithstanding the limitations imposed by Regulation Fair Disclosure, being able to communicate with managers will improve their trading decisions. For investment managers following active strategies, trading decisions that change the weight of a portfolio company relative to its weighting in the index are likely to be the main determinants of their performance relative to their benchmark index. To the extent that active

investment managers believe that making stewardship decisions that corporate managers disfavor might adversely affect their access to such managers, an interest in improving relative performance could provide incentives to avoid such decisions.

Note that, to the extent that investment managers get access to corporate managers and consequently make better trading decisions, the gains from such trading decisions will improve the investment manager's performance relative to others, since rivals will not share these trading gains. By comparison, gains from governance-generated improvements in the value of particular portfolio companies will be substantially shared with rivals. Thus, an interest in improving relative performance could well lead active fund managers to place more weight on gains to their portfolios from access to corporate managers relative to gains from governance-generated increases in value, compared to what would be optimal for the investment funds' beneficial investors.[3]

Finally, without discussing the issue in detail, we want to flag a disagreement in the literature regarding the extent to which fund inflows and outflows are sensitive to changes in relative performance (for example, Sirri and Tufano 1998 and Coates and Hubbard 2007). To the extent that the sensitivity of inflows and outflows to performance is limited, competition with other investment funds will give investment managers limited incentives to improve the value of portfolio companies.

The Governance Passivity of Investment Funds

The above analysis suggests that investment managers, those managing both passive index funds and active mutual funds, have incentives to be "more passive" with respect to governance issues than is optimal for their beneficial investors.

With respect to index funds, our analysis is consistent with the practically negligible resources that index funds spend on stewardship beyond what is required to comply with regulations requiring investment managers to vote shares in portfolio companies and to avoid doing so in an uninformed fashion. Vanguard employs about 15 staff for voting and stewardship at its 13,000 portfolio companies; BlackRock employs 24 staff for voting and stewardship at 14,000 portfolio companies; and State Street Global Advisors employs fewer than 10 staff for voting and stewardship at 9,000 portfolio companies (Krouse, Benoit, and McGinty 2016).

Of course, these staff may receive information from proxy advisors as well as from active portfolio managers employed by the investment manager. However, each of these major investment managers devotes less than one person-workday per year, on average, to assessing this and other information, and undertaking other stewardship activities with respect to each of their portfolio companies. Note that each of these investment managers is likely to hold several percent of each company's stock and to be among their largest shareholders. Given the size and value of the positions

[3] An increase in relative gross returns could be used by an investment manager not to attract additional funds but to extract an increase in the level of fees charged without risking an outflow of funds. The above analysis, suggesting that an interest in increasing relative performance is unlikely to induce optimal stewardship decisions, also applies equally to this scenario.

that each of these investment managers holds in large public companies, there are grounds for concern that these managers substantially underinvest in stewardship.

With respect to active mutual funds, our analysis is similarly consistent with the very limited resources that predominantly actively managed mutual fund families currently spend on stewardship. Even the largest such mutual fund families employ only a small number of staff to make voting decisions and undertake all other governance-related stewardship activities in the vast number of corporations in which they hold stock.

In a companion paper, we document that this underinvestment by investment managers is reflected not only in the limited time that their staff spend on voting and stewardship activities, but also in the absence of these investment managers from the ranks of investors that use certain significant tools to generate value increases from improved governance that benefit the investment funds (Bebchuk, Cohen, and Hirst 2017). For example, large investment managers generally avoid submitting shareholder proposals, nominating directors to the boards of corporations, or conducting proxy contests. Their absence might be due not only to incentives to underspend on stewardship, but also to private costs that investment managers viewed as oppositional to managers might have to bear, which we discuss below.

Our companion paper also addresses the argument that substantial passivity on the part of investment managers is optimal, and that the underspending problem is therefore of limited economic importance. Such an argument could be justified if other mechanisms—such as the discipline of the market for corporate control, executive incentives schemes, or monitoring and engagement by other investors—could be relied on to eliminate agency problems in public companies. We argue, however, that the limits of such mechanisms make it plausible to assume that improved stewardship by the investment managers that hold a large proportion of the shares of most publicly traded companies can significantly improve outcomes for their own investors.

There is a growing recognition of the power of large investment managers, and concomitantly increasing expectation that they will use this power to improve the governance of their portfolio companies. The leaders of the largest index fund managers have responded by making public announcements stressing their commitment to stewardship, and to improving corporate governance (for example, Fink 2015; McNabb 2015). These executives may indeed believe in the desirability of governance improvements and sincerely wish to help bring them about. However, our economic analysis indicates that investment managers may well have very limited economic incentives to spend on stewardship, and may have economic incentives to be more lax toward corporate managers, compared with what would be optimal for their beneficial investors.

Private Costs from Opposing Managers

Another significant source of agency problems introduced by the separation between investment managers and beneficial owners is that investment managers may bear private costs from taking positions that corporate managers disfavor. When

such private costs may result, investment managers may be more reluctant to spend on actions or make qualitative decisions that are disfavored by corporate managers. Suppose that such an action would result in a change in the value of the portfolio of ΔV but a private indirect cost of IC to the investment manager. The investment manager will take the disfavored action only if $C + IC$ is less than $\alpha \times \Delta V$.

For qualitative choices that would not involve any additional marginal cost but would have an expected positive effect on the value of the portfolio (that is, $\Delta V > 0$), the investment manager would prefer to side with managers if $IC > \alpha \times \Delta V$. Thus, the investment manager would prefer to avoid taking a position disfavored by managers that would be optimal for the managed portfolio if and only if:

$$0 < \Delta V < \frac{IC}{\alpha}.$$

What is important is not whether avoiding such actions actually helps investment managers obtain business, but whether investment managers believe that to be the case, on an expected value basis. The smaller is α, the wider the range of increases in value that the investment manager would forgo not to bear expected indirect costs of taking actions that corporate managers disfavor. That investment funds charge fees below 1 percent (on average) strengthens the distortion resulting from potential indirect costs.

One important source of costs from taking positions that corporate managers disfavor (or benefits from taking positions that managers favor) comes from the incentives of investment managers to obtain or retain business from public corporations. In 2015, 401(k) assets under management totaled $4.7 trillion, with 60 percent held in mutual funds (Collins, Holden, Duvall, and Chism 2016, p. 2); most of these assets are likely to come from public corporations. Cvijanović, Dasgupta, and Zachariadis (2016) document that an average of 14 percent of fund family revenue is derived from 401(k)-related business. The largest index fund managers and active managers all derive business from 401(k) services, and therefore have strong incentives to attract and retain such business from public corporations.

In addition, many investment managers provide investment services to corporations, both to manage cash and short-term investments and also to manage the long-term investments of financial corporations such as insurance companies. Investment managers may also provide investment management services to pension funds that are sponsored by public corporations, and over which the corporation may have some influence. US private sector pension funds had aggregate assets under management of $2.9 trillion in 2015 (Investment Company Institute 2016). Several empirical studies provide evidence suggesting that business ties with corporations influence the voting decisions of investment managers. Davis and Kim (2007) find that the volume of pension fund business of investment managers was associated with those investment managers voting more often with corporate managers on several key types of shareholder proposals. Ashraf, Jayaraman, and Ryan (2012) find that mutual fund families that have greater business ties to corporations tend to vote

more favorably toward corporate managers on executive compensation matters at all corporations.

These studies focus on the association between corporate business ties in general and voting in corporations in general. Cvijanović, Dasgupta, and Zachariadis (2016) examine contested shareholder proposals where corporate managers care more about votes for their favored position, and find that mutual fund families with business ties to a corporation are more likely to cast pro-management votes in closely contested situations at the corporation. Although this study provides evidence that an investment manager's business ties with particular corporations provide incentives to vote with corporate managers in close votes, there are clear limits to the ability of investment managers to treat managers of client corporations more favorably than their general voting policy would provide. Therefore, in our view, the more important concern is that investment managers will have an incentive to lean in a pro-management direction when determining their strategies and policies regarding stewardship.

Given the limited economic incentive that investment managers have to generate governance gains in portfolio companies, and their strong economic interest in attracting more business, choosing a pro-management approach within the range of the legitimate choices available to them may seem the safest approach to investment managers. Investment managers would have an incentive to take such an approach as long as they believe that doing so might help them get additional business from public corporations on an expected value basis.

Finally, we note certain additional private costs that are relevant only to the largest investment managers and may contribute to discouraging these major players from opposing corporate managers. Some mutual fund families hold close to or above 5 percent of the stock in many public corporations. Indeed, the three index fund managers that dominate the index fund sector—Vanguard, BlackRock, and State Street Global Advisors—hold such positions in most large publicly traded corporations; Fidelity Investments and the Capital Group also hold such positions in many public corporations, and Dimensional Fund Advisors holds such positions in many smaller public corporations. Investment managers holding such positions would bear additional private costs in the event that they attempt to wield significant influence—and therefore have a significant incentive to avoid doing so.

Under Section 13(d) of the Exchange Act of 1934, investors that own or control, in the aggregate, 5 percent or more of a corporation's shares and that seek to influence the control of the corporation are subject to extensive and repeated disclosure requirements on Schedule 13D. Nominating directors, undertaking a proxy contest for board representation, and other significant engagement action would classify investment managers as seeking to influence control. By contrast, investment managers that are not classified as seeking to influence control are subject only to the relatively limited disclosure requirements on Schedule 13G. Becoming subject to the substantial and repeated disclosure on Schedule 13D would be very costly for the investment managers of major fund families, which

typically manage multiple funds. Because the investment manager would have to bear these costs itself rather than charge them to the investment funds, the prospect of having to bear such costs provides additional incentives to avoid taking any actions that might be classified as seeking to influence the control of the corporation.

Activist Hedge Funds

Finally, we would like to discuss a different type of an investment manager, the activist hedge fund manager. Applying the framework described above shows why activist hedge fund managers suffer less from the agency problems that affect investment managers with diversified equity portfolios, and why activist hedge fund managers have incentives to make stewardship decisions that are significantly closer to those that would be optimal for their beneficial investors.

Why Activist Hedge Funds are Different

Hedge funds managers limit their investment offerings to investors considered to be sophisticated, and are therefore not subject to the regulations governing investment managers of mutual funds. Hedge funds therefore have considerably more freedom in the assets they own, their use of leverage, and their compensation structures. Our focus below is on the subset of hedge funds that take concentrated positions in the equity of public corporations and actively engage with corporate managers—activist hedge funds. For the reasons explained below, these hedge funds have significant influence on the corporate governance landscape.

High-Powered Incentives to Increase Value. Hedge fund managers, including activist hedge fund managers, typically receive compensation based on two components, often referred to as "2 and 20" (French 2008): a management fee that is a relatively small percentage of the value of the assets, historically 2 percent, and an incentive payment, structured as a "carried interest" of a proportion (historically 20 percent) of any increase in value of the portfolio.

Leaving aside the management fee, which is higher than the average for an actively managed mutual fund but a similar order of magnitude, a hedge fund manager that is able to increase the value of a position in a portfolio company through investments in stewardship will capture 20 percent of this increase, an order of magnitude more than the percentage of any value increase that a mutual fund manager would be able to capture. Thus, activist hedge fund managers will have much stronger incentives to bring about governance-generated increases in value than investment managers of mutual funds, even when the latter hold positions with equal or greater dollar value.

Limited Business from Portfolio Companies. In contrast to mutual funds, which are registered investment companies and publicly issue securities, hedge funds are not registered investment companies and do not accept investments from 401(k) plans. Accordingly, activist hedge fund managers do not have a desire to attract

401(k) business that might discourage them from taking positions that corporate managers disfavor. In addition, activist hedge funds do not offer other services to corporations of the kind that many investment managers offer.

Concentrated Positions and Stronger Incentives Regarding Relative Performance. Activist hedge funds have concentrated positions, sometimes holding significant positions in as few as 10 portfolio companies. As a result, an improvement in the value of a single portfolio company that is a target of stewardship activities can substantially improve the fund manager's performance relative to peer investment vehicles. This will, in turn, affect the manager's ability to attract additional investments. For example, the investment of Pershing Square Capital Management LP in Canadian Pacific Railway Ltd. and General Growth Properties Inc. each constituted as much as one-fifth of the fund's portfolio during certain periods, and the increase in the value of these positions enabled the fund to post strong performance.

Because of their small size and method of selection, activist hedge fund portfolios display very little correlation with those of competing funds, or with other investment opportunities available to their investors. Any changes in the value of their portfolio companies are therefore also clearly reflected in their relative performance against such comparable investments. This factor therefore strengthens the incentive of activist hedge fund managers to bring about governance-related improvements in the value of their portfolio companies. Thus, the desire to improve relative performance provides more powerful incentives for activist hedge funds to seek governance-related value improvements than it does for managers of index funds and active mutual funds.

Clearly, the main factors that create a wedge between the interests of investment managers and the beneficial investors whose investments they manage affects activist hedge fund managers significantly less than investment managers of mutual funds. Consistent with this, activist hedge fund managers are much more willing to devote significant resources to stewardship. Activist hedge fund managers are often willing to devote hundreds of person-hours per year to monitoring and engaging with each of their portfolio companies. For instance, Pershing Square Capital Management has an investment team of eight, plus several other employees, that oversee a portfolio of about 12 corporations (as reported in Krouse, Benoit, and McGinty 2016). Activist hedge fund managers are also willing to have representatives on the board of directors of portfolio companies, and often seek such representation. Such representation not only requires significant personnel time, but also imposes constraints on the activist hedge fund manager's trading in the portfolio company's stock.

Furthermore, activist hedge fund managers frequently commence proxy contests at portfolio companies (Brav, Jiang, Partnoy, and Thomas 2008), despite the considerable expenses associated with such contests (estimated by Gantchev 2013 to average about $10 million) and corporate managers' views of such contests as adversarial. By contrast, managers of mutual funds have generally avoided conducting proxy contests at their portfolio companies, even where the mutual fund held a significant stake. Even in situations where activist hedge fund managers do not conduct proxy contests, they frequently take public positions that the managers of their portfolio companies disfavor, which other investment managers generally avoid.

Clearly activist hedge fund managers have different incentive structures that enable them to play an important role in the current governance landscape. This role is especially important in light of the significant agency problems that afflict the stewardship decisions of mutual fund managers. But while activist hedge fund managers play a beneficial role in the corporate governance system, there are significant limits to this beneficial role.

The Limits of Hedge Funds

Activist hedge fund managers have incentives to spend on stewardship only when the governance-generated value increases likely to result are especially large. The incentives of activist hedge fund managers are driven by the significant performance-related fees that they earn, and by their concentrated portfolios. As a result, activist hedge fund managers can pursue only those corporations where the potential governance-related increases in value are sufficiently large that the funds' investors can expect to make reasonable risk-adjusted returns after bearing the high fees charged by the hedge fund managers and the firm-specific risks from the funds' concentrated portfolios. For example, where an activist hedge fund could buy a stake in a given corporation and bring about a 3 percent increase in value over a two-year period, the hedge fund manager would be unlikely to pursue this opportunity.

This analysis is consistent with the fact that such funds usually focus on situations where governance failures have led to substantial operating underperformance. As a result, disclosures regarding the initiation of engagements by activist hedge fund managers are accompanied by abnormal returns that, on average, exceed 5 percent, reflecting market expectations of a significant expected increase in value (for example, Brav et al. 2008; Bebchuk, Brav, and Jiang 2015).

Furthermore, for an activist hedge fund manager to bring about a governance-generated increase in value, it is not only necessary that there be potential for such a large increase, but also that other institutional investors are willing to support the changes sought by the activist hedge fund manager. Activist hedge fund managers are unable to bring about changes unless they obtain the support of other types of institutional investors, or have a reasonable likelihood of doing so (Bebchuk and Jackson 2012). When an activist hedge fund manager enjoys such support for the changes it seeks, it will be able to win a proxy fight, or obtain a settlement by credibly threatening to do so, and thereby cause the corporation to make such changes. Conversely, when corporate managers expect that most institutional investors will side with them and not with activist hedge fund managers, activist hedge fund managers will not have much influence.

Mutual fund managers do sometimes vote on the side of activist hedge fund managers. Indeed, the expectation that this would be the case, and that activist hedge funds could therefore prevail in potential proxy fights, often leads corporate managers to accept activist hedge funds' demands for board representation (Bebchuk, Brav, Jiang, and Keusch 2017). However, our analytical framework raises the concern that, on the margin, mutual fund managers might not be sufficiently

willing to support activist hedge fund managers in their engagements with portfolio companies where such support would be optimal for the mutual funds' investors. Whether and to what extent this is the case is an interesting issue for future research.

Finally, we should briefly note the issue of short-termism and long-termism. Activist hedge fund managers have stronger incentives to bring about increases in value than other institutional investors. However, some scholars have argued that activist hedge fund managers focus on increases in short-term value and that the increases they seek often come at the expense of long-term value (for example, Strine 2014; Coffee and Palia 2015). One of us has addressed this claim in detail elsewhere on both conceptual and empirical grounds (Bebchuk 2013; Bebchuk, Brav, and Jiang 2015). Leaving aside the alleged distinction between short-term and long-term increases in value, a key point of our analysis is that activist hedge fund managers stand out relative to other institutional investors in terms of their incentives to seek increased value.

Of course, index funds are long-term players, and can therefore be expected to favor only changes that would enhance value in the long term (for examples of this view, see Lipton 2014, 2016). But our analysis shows that investment managers overseeing index funds have very limited incentives to bring about governance-generated increases in value, be they long-term or short-term.

Implications

The rise of institutional investors has transformed the governance landscape facing the modern corporation. With shares concentrated in the hands of institutional investors, corporate managers no longer face diffuse shareholders that are powerless to engage with managers. However, the agency problems of institutional investors prevent the full realization of the potential benefits of the increased concentration of shareholdings. Investment managers overseeing diversified equity portfolios have incentives to spend considerably less on stewardship, and to side with corporate managers more frequently, than would be optimal for their beneficial investors. These factors operate to suppress investor stewardship relative to optimal levels.

In this paper, we have provided a framework for analyzing these agency problems. We have also applied this framework to several key categories of investment managers. Our analysis has significant implications for researchers and policymakers. While a full analysis of these implications is beyond the scope of this paper, we outline ten of these implications below.

1. *Research.* Over recent decades, the amount of academic work analyzing agency problems in corporate governance has increased dramatically (for example, Bebchuk, Cohen, and Wang 2013), but most of this work has examined the agency problems of corporate insiders. We hope that our work will stimulate and provide a framework for future work on the agency problems of institutional investors.

2. *Disclosure.* Public awareness and academic research about the agency problems of managers of publicly traded corporations is facilitated by the extensive disclosures made by such corporations about internal decisions. Policymakers may wish to consider adopting regulations that would require investment managers to disclose information that would enable investors and others to identify and assess agency problems. For example, investment managers of mutual funds have been required to disclose how they vote their shares in publicly traded corporations since 2004, but some other investment managers are not required to do so. Furthermore, policymakers may want to consider tighter disclosure requirements that would provide comprehensive information about the business ties between investment managers and the public corporations in which they invest.

3. *Regulation of Mutual Fund Fees.* Regulations that preclude key investment managers from charging stewardship expenses to their investment funds, or from tying fees to increases in the value of their portfolios, have significant effects on the stewardship decisions of these investment managers. These regulations might be justified to protect the beneficial investors in these investment funds. However, policymakers should recognize the tradeoffs created by these rules, and consider whether some adjustments may be warranted.

4. *Stewardship Codes.* In a number of countries, such as the United Kingdom (Financial Reporting Council 2012), Japan (Council of Experts Concerning the Japanese Version of the Stewardship Code 2014), and Canada (Office of the Superintendent of Financial Institutions Canada 2013), concerns about whether institutional investors undertake adequate stewardship have led to the development of nonbinding stewardships codes which various institutional investors have pledged to follow. Our analysis suggests that there is a problem with the incentives of institutional investors to spend on stewardship. To the extent that this is the case, stewardship codes putting forward aspirations, principles, or guidelines are likely to have less of an impact than if investment managers had appropriate incentives.

5. *Index Investing.* The rise of index investing has generally been viewed as a positive development because it has reduced the cost of investment intermediation. Our analysis shows that a continuation of this trend could have significant costs for corporate governance. This analysis also highlights the challenges likely to result if index funds continue to grow as expected.

6. *Anticompetitive Effects of Index Investing.* Recent work has raised concerns that, because index funds are invested across various corporations in an economic sector, they would have incentives to encourage those corporations to engage in anticompetitive behavior that would enable them to capture monopolistic rents, (for example, Elhauge 2016; Posner, Scott Morton, and Weyl 2016).[4] This line of work is based on the premise that index fund managers have strong incentives to take whatever actions would maximize the collective wealth of their beneficial investors. Our analysis indicates that index fund managers might well have different incentives,

[4]These arguments build on empirical studies by Azar, Schmalz, and Tecu (2017) and Azar, Raina, and Schmalz (2016), although these studies have recently been questioned by Rock and Rubinfeld (2017).

which would lead them to limit intervention with their portfolio companies. Thus, our analysis suggests that it is implausible to expect that index fund managers would seek to facilitate significant anticompetitive behavior.

7. *Proxy Advisors.* Institutional investors commonly employ the services of one or more proxy advisors, such as ISS and Glass Lewis, which analyze voting choices faced by investors in public corporations and make recommendations (Malenko and Shen 2016). Critics of proxy advisors would prefer that institutional investors reduce their reliance on the analysis and recommendations provided by proxy advisors (Clark and Van Buren 2013). Indeed, legislation currently being considered by Congress (previously titled the Proxy Advisory Firm Reform Act of 2016) would regulate proxy advisors in ways that might significantly increase their costs of operation and otherwise discourage their activities. Our analysis raises a concern that a reduction in the activities of proxy advisors would not be offset by increased spending on analysis by institutional investors sufficient to maintain even their current levels of monitoring.

8. *Hedge Fund Activism.* There is a heated debate over the role of hedge fund activism. Whereas some writers, including one of us, have been supportive of such activism (for example, Bebchuk and Jackson 2012; Bebchuk 2013; Bebchuk, Brav, and Jiang 2015; Gilson and Gordon 2013), others view it as counterproductive and advocate various measures that would limit and discourage such activism (for example, Strine 2014; Coffee and Palia 2015). Some prominent critics of hedge fund activism would like to see the engagement currently conducted by activist hedge fund managers replaced by the stewardship of institutional investors. Our analysis shows the important role that activist hedge fund managers play in the corporate governance landscape. Because the incentives of mutual fund managers differ substantially from those of activist hedge fund managers, were the abilities of hedge funds to undertake such engagement to be impeded, stewardship by mutual fund managers would be unlikely to replace activist hedge fund managers in constraining agency problems in public corporations.

9. *Wolf Packs.* When an activist hedge fund takes a position in an underperforming public corporation, other hedge funds often acquire positions in the corporation (Brav, Dasgupta, and Mathews 2016). Groups of such "follower" hedge funds are commonly referred to as "wolf packs," and various writers have suggested that they are a negative influence (for example, Coffee and Palia 2015). Our analysis, however, indicates that so-called wolf packs might serve a useful purpose. Because mutual funds might be reluctant to vote against incumbents, an activist hedge fund might sometimes be unable to win a proxy fight against underperforming incumbents when such victory would be in the interests of investors. By contrast, when a dispute between incumbents and an activist hedge fund draws other hedge funds to invest, the new shareholders are more willing to also invest in assessing which course of action would be optimal and to vote accordingly, including voting against the incumbents if they conclude it to be value-enhancing.

10. *Shareholder Rights.* For some critics of shareholder rights (Bainbridge 2006, for example), the imperfections of institutional investors, and the fact that

stewardship decisions are taken by agents rather than the ultimate beneficial investors, provide a rationale for weakening shareholder rights and insulating corporate managers from shareholder action. Given that the agents may not be acting in the interests of beneficial investors, so the argument goes, there is reason to limit the power of the tools given to those agents lest they use the tools in ways that are counterproductive to the interests of their beneficial investors. However, our analysis of the agency problems of institutional investors identifies a clear direction in which their stewardship decisions deviate from those that are optimal for their beneficial investors: investment managers can be expected to underutilize the tools they have to engage with corporate managers.

Thus, notwithstanding the imperfections of investment managers as agents for their beneficial investors, there is little basis for concerns that institutional investors will interfere excessively with the actions of corporate managers. Accordingly, there is no reason to weaken shareholder rights or impede shareholder action based on such concerns. An understanding of the agency problems of institutional investors leads to the conclusion that modern corporations do not suffer from too much shareholder intervention, but rather from too little.

■ *The authors are grateful to Matthew Bodie, Robert Clark, John Coates, Stephen Davis, Einer Elhauge, Jesse Fried, Mark Gertler, Gordon Hanson, Kobi Kastiel, Reinier Kraakman, Enrico Moretti, Mark Roe, Robert Sitkoff, Holger Spamann, Timothy Taylor, and participants in three Harvard seminars and the American Law and Economics Association Annual Meeting for valuable suggestions; to David Mao and Gregory Shill for excellent research assistance; and to Harvard Law School for financial support.*

References

Admati, Anat R., and Paul Pfleiderer. 2009. "The 'Wall Street Walk' and Shareholder Activism: Exit as a Form of Voice." *Review of Financial Studies* 22(7): 2645–85.

Aghion, Philippe, John Van Reenen, and Luigi Zingales. 2013. "Innovation and Institutional Ownership." *American Economic Review* 103(1): 277–304.

Ashraf, Rasha, Narayanan Jayaraman, and Harley E. Ryan Jr. 2012. "Do Pension-Related Business Ties Influence Mutual Fund Proxy Voting? Evidence from Shareholder Proposals on Executive Compensation." *Journal of Financial and Quantitative Analysis* 47(3): 567–88.

Azar, José, Sahil Raina, and Martin C. Schmalz. 2016. "Ultimate Ownership and Bank Competition." Available at SSRN: https://papers.ssrn.com/abstract=2710252.

Azar, José, Martin C. Schmalz, and Isabel Tecu. 2017. "Anti-Competitive Effects of Common Ownership." Available at SSRN: https://papers.ssrn.com/abstract=2427345.

Bainbridge, Stephen M. 2006. "Director Primacy and Shareholder Disempowerment." *Harvard Law Review* 119(6): 1735–58.

Bebchuk, Lucian A. 2013. "The Myth that

Insulating Boards Serves Long-Term Value Essay." *Columbia Law Review* 113(6): 1637–94.

Bebchuk, Lucian A., Alon Brav, and Wei Jiang. 2015. "The Long-Term Effects of Hedge Fund Activism." *Columbia Law Review* 115(5): 1085–155.

Bebchuk, Lucian A., Alon Brav, Wei Jiang, and Thomas Keusch. 2017. "Dancing with Activists." Available at SSRN: https://papers.ssrn.com/abstract=2948869.

Bebchuk, Lucian A., Alma Cohen, and Scott Hirst. 2017. "The Under-Supply of Shareholder Engagement." Unpublished.

Bebchuk, Lucian A., Alma Cohen, and Charles C. Y. Wang. 2013. "Learning and the Disappearing Association between Governance and Returns." *Journal of Financial Economics* 108(2): 323–48.

Bebchuk, Lucian A., and Robert J. Jackson Jr. 2012. "The Law and Economics of Blockholder Disclosure." *Harvard Business Law Review* 2: 39–60.

Bebchuk, Lucian A., and Zvika Neeman. 2010. "Investor Protection and Interest Group Politics." *Review of Financial Studies* 23(3): 1089–119.

Berle, Adolf A., and Gardiner C. Means. 1932. *The Modern Corporation and Private Property*. New York: Macmillan.

Board of Governors of the Federal Reserve System. 2016. "Financial Accounts of the United States: Flow of Funds, Balance Sheets, and Integrated Macroeconomic Accounts, Z.1." https://www.federalreserve.gov/releases/z1/current/ (accessed April 15, 2017).

Brav, Alon, Amil Dasgupta, and Richmond D. Mathews. 2016. "Wolf Pack Activism." Available at SSRN: https://papers.ssrn.com/abstract=2840704.

Brav, Alon, Wei Jiang, Frank Partnoy, and Randall Thomas. 2008. "Hedge Fund Activism, Corporate Governance, and Firm Performance." *Journal of Finance* 63(4): 1729–75.

Clark, Cynthia E., and Harry J. Bus Van Buren. 2013. "Compound Conflicts of Interest in the US Proxy System." *Journal of Business Ethics* 116(2): 355–71.

Coates, John C. IV, and R. Glenn Hubbard. 2007. "Competition in the Mutual Fund Industry: Evidence and Implications for Policy." *Journal of Corporation Law* 33(1): 151–222.

Coffee, John C. Jr., and Darius Palia. 2015. *The Wolf at the Door: The Impact of Hedge Fund Activism on Corporate Governance*. Boston: Now Publishers Inc.

Collins, Sean, Sarah Holden, James Duvall, and Elena Barone Chism. 2016. "The Economics of Providing 401(k) Plans: Services, Fees, and Expenses, 2015." *ICI Research Perspective* 22(4): 1–31.

Council of Experts Concerning the Japanese Version of the Stewardship Code. 2014. *Principles for Responsible Institutional Investors*. Tokyo, Japan: Financial Services Agency.

Cremers, K. J. Martijn, and Antti Petajisto. 2009. "How Active Is Your Fund Manager? A New Measure that Predicts Performance." *Review of Financial Studies* 22(9): 3329–65.

Cvijanović, Dragana, Amil Dasgupta, and Konstantinos E. Zachariadis. 2016. "Ties that Bind: How Business Connections Affect Mutual Fund Activism." *Journal of Finance* 71(6): 2933–66.

Davis, Gerald F., and E. Han Kim. 2007. "Business Ties and Proxy Voting by Mutual Funds." *Journal of Financial Economics* 85(2): 552–70.

Demsetz, Harold. 1983. "The Structure of Ownership and the Theory of the Firm." *Journal of Law and Economics* 26(2): 375–90.

Diamond, Randy. 2016. "Established Firms Keep Lion's Share of Business." *Pensions and Investments*, November 14. http://www.pionline.com/article/20161114/PRINT/311149993/established-firms-keep-lions-share-of-business.

Elhauge, Einer. 2016. "Horizontal Shareholding." *Harvard Law Review* 129(5): 1267–317.

FactSet Research Systems. No date. FactSet Ownership database. https://www.factset.com/data/company_data/ownership.

Financial Reporting Council. 2012. *UK Stewardship Code*. London: Financial Reporting Council.

Fink, Laurence D. 2015. *Text of Letter Sent to Chairmen/CEOs Asking Them to Focus on Delivering Long-Term Value*. New York: BlackRock.

French, Kenneth R. 2008. "Presidential Address: The Cost of Active Investing." *Journal of Finance* 63(4): 1537–73.

Gantchev, Nickolay. 2013. "The Costs of Shareholder Activism: Evidence from a Sequential Decision Model." *Journal of Financial Economics* 107(3): 610–31.

Gilson, Ronald J., and Jeffrey N. Gordon. 2013. "The Agency Costs of Agency Capitalism: Activist Investors and the Revaluation of Governance Rights." *Columbia Law Review* 113(4): 863–928.

Grossman, Sanford J., and Oliver D. Hart. 1980. "Takeover Bids, The Free Rider Problem, and the Theory of the Corporation." *Bell Journal of Economics* 11(1): 42–64.

Hartzell, Jay C., and Laura T. Starks. 2003. "Institutional Investors and Executive Compensation." *Journal of Finance* 58(6): 2351–74.

Hirst, Scott. 2016. "Social Responsibility Resolutions." Available at SSRN: https://papers.ssrn.com/abstract=2773367.

Investment Company Institute. 2016. *2016 Investment Company Fact Book*. Washington, DC: Investment Company Institute.

Ke, Bin, Kathy R. Petroni, and Yong Yu. 2008. "The Effect of Regulation FD on Transient

Institutional Investors' Trading Behavior." *Journal of Accounting Research* 46(4): 853–83.

Krouse, Sarah, David Benoit, and Tom McGinty. 2016. "Meet the New Corporate Power Brokers: Passive Investors." *Wall Street Journal*, October 24.

Lipton, Martin. 2014. "Current Thoughts about Activism, Revisited." *Harvard Law School Forum on Corporate Governance and Financial Regulation*, April 8. https://corpgov.law.harvard.edu/2014/04/08/current-thoughts-about-activism-revisited/.

Lipton, Martin. 2016. "The New Paradigm for Corporate Governance." *Harvard Law School Forum on Corporate Governance and Financial Regulation*, February 3. https://corpgov.law.harvard.edu/2016/02/03/the-new-paradigm-for-corporate-governance/.

Malenko, Nadya, and Yao Shen. 2016. "The Role of Proxy Advisory Firms: Evidence from a Regression-Discontinuity Design." *Review of Financial Studies* 29(12): 3394–427.

Manne, H. G. 1965. "Mergers and the Market for Corporate Control." *Journal of Political Economy* 73(2): 110–20.

McNabb, Bill. 2015. "Investors: Getting to Know You and Your Governance." *The Corporate Board*, March–April.

Oey, Patricia, and Christina West. 2016. *Average Fund Costs Continued to Decline in 2015 But Investors Are Not Necessarily Paying Less*. Morningstar Manager Research.

Office of the Superintendent of Financial Institutions Canada. 2013. *Corporate Governance Guideline*. Ottawa: Office of the Superintendent of Financial Institutions Canada.

Posner, Eric A., Fiona M. Scott Morton, and E. Glen Weyl. 2016. "A Proposal to Limit the Anti-Competitive Power of Institutional Investors." Available at SSRN: https://papers.ssrn.com/abstract=2872754.

Rock, Edward B., and Daniel L. Rubinfeld. 2017. "Defusing the Antitrust Threat to Institutional Investor Involvement in Corporate Governance." Available at SSRN: https://papers.ssrn.com/abstract=2925855.

Shleifer, Andrei, and Robert W. Vishny. 1986. "Large Shareholders and Corporate Control." *Journal of Political Economy* 94(3): 461–88.

Sirri, Erik R., and Peter Tufano. 1998. "Costly Search and Mutual Fund Flows." *Journal of Finance* 53(5): 1589–622.

Strine, Leo E. Jr. 2014. "Can We Do Better by Ordinary Investors? A Pragmatic Reaction to the Dueling Ideological Mythologists of Corporate Law Essay." *Columbia Law Review* 114(2): 449–502.

Tergesen, Anne, and Jason Zweig. 2016. "The Dying Business of Picking Stocks." *Wall Street Journal*, October 17. http://www.wsj.com/articles/the-dying-business-of-picking-stocks-1476714749.

Tonello, Matteo, and Stephan Rahim Rabimov. 2010. "The 2010 Institutional Investment Report: Trends in Asset Allocation and Portfolio Composition." Available at SSRN: https://papers.ssrn.com/abstract=1707512.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 113–130

Towards a Political Theory of the Firm

Luigi Zingales

The revenues of large companies often rival those of national governments. In a list combining both corporate and government revenues for 2015, ten companies appear in the largest 30 entities in the world: Walmart (#9), State Grid Corporation of China (#15), China National Petroleum (#15), Sinopec Group (#16), Royal Dutch Shell (#18), Exxon Mobil (#21), Volkswagen (#22), Toyota Motor (#23), Apple (#25), and BP (#27) (Global Justice Now 2016). All ten of these companies had annual revenue higher than the governments of Switzerland, Norway, and Russia in 2015. Indeed, 69 of the largest 100 corporate and government entities ranked by revenues were corporations. In some cases, these large corporations had private security forces that rivaled the best secret services, public relations offices that dwarfed a US presidential campaign headquarters, more lawyers than the US Justice Department, and enough money to capture (through campaign donations, lobbying, and even explicit bribes) a majority of the elected representatives. The only powers these large corporations missed were the power to wage war and the legal power of detaining people, although their political influence was sufficiently large that many would argue that, at least in certain settings, large corporations can exercise those powers by proxy.

Yet in contemporary economics, the commonly prevailing view of the firm ignores all these elements of politics and power. According to this view, the firm is a simple "nexus of contracts" (Jensen and Meckling 1976), with no objectives or

■ *Luigi Zingales is the Robert C. McCormack Distinguished Service Professor of Entrepreneurship and Finance, University of Chicago Booth School of Business, Chicago, Illinois. His email address is Luigi.Zingales@ChicagoBooth.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.113

doi=10.1257/jep.31.3.113

life separate from those of its contracting parties, a veil or—better—a handy tool for individuals to achieve their personal goals. This view might be a reasonable approximation for small or closely held private corporations, but certainly it does not accurately describe giant global corporations.

The largest modern corporations facilitated a massive concentration of economic (and political) power in the hands of a few people, who are hardly accountable to anyone. The reason is not simply that many of those giants (like State Grid, China National Petroleum, and Sinopec) are overseen by members of the Chinese Communist party. In the United States, hostile takeovers of large corporations have (unfortunately) all but disappeared, and corporate board members are essentially accountable to no one. Only rarely are they not re-elected, and even when they do not get a plurality of votes, they are co-opted back to the very same board (Committee on Capital Market Regulation 2014). The primary way for board members to lose their jobs is to criticize the incumbent chief executive officer (see the Bob Monks experience in Tyco described in Zingales 2012). The only genuine pressure on large US corporations from the marketplace is exercised by activist investors, who operate under strong political opposition and not always with the interests of all shareholders in mind.

In this essay, I will argue that the interaction of concentrated corporate power and politics is a threat to the functioning of the free market economy and to the economic prosperity it can generate, and a threat to democracy as well. I begin with a discussion of how these concerns were present in Adam Smith's (1776) work, how they were neutralized in the neoclassical theory of the firm, and then how they were reborn, at least to a certain extent, in the "incomplete contracts view" of the firm. However, even the incomplete contracts view is designed for an environment in which the rules of the game are exogenously specified and enforced. Once we recognize, however, that large firms have considerable power in influencing the rules of the game, important questions arise: To what extent can the power firms have in the marketplace be transformed into political power? To what extent can the political power achieved by firms be used to protect but also enhance the market power firms have?

The phenomenon of corporations becoming large enough to influence and in some cases to dominate politics is not new. In their 1932 classic, *The Modern Corporation and Private Property*, Berle and Means wrote: "The rise of the modern corporation has brought a concentration of economic power which can compete on equal terms with the modern state—economic power versus political power, each strong in its own field. The state seeks in some aspects to regulate the corporation, while the corporation, steadily becoming more powerful, makes every effort to avoid such regulation. ... The future may see the economic organism, now typified by the corporation, not only on an equal plane with the state, but possibly even superseding it as the dominant form of social organization."

I will argue that US economic patterns in the last few decades have seen a rise in the relative size of large companies. Thus, I call attention to the risk of a "Medici vicious circle," in which economic and political power reinforce each other.

The "signorias" of the Middle Ages—the city-states that were a common form of government in Italy from the 13th through the 16th centuries—were a takeover of a democratic institution ("communes") by rich and powerful families who ran the city-states with their own commercial interests as a main objective. The possibility and extent of this Medici vicious circle depend upon several nonmarket factors. I identify six of them: the main source of political power, the conditions of the media market, the independence of the prosecutorial and judiciary power, the campaign financing laws, and the dominant ideology. I describe when and how these factors play a role and how they should be incorporated in a broader "political theory" of the firm.

From Adam Smith to the Neoclassical Theory of the Firm

Adam Smith's View of Joint Stock Companies

Economists have not always been blind to the power of corporations. Adam Smith (1776 [1904], Book V, chap. 1) himself had a very negative view of corporations, then called joint stock companies, which were granted monopoly rights by the Crown: "The directors of such companies, however, being the managers rather of other people's money than of their own, it cannot well be expected that they should watch over it with the same anxious vigilance with which the partners in a private copartnery frequently watch over their own. … Negligence and profusion, therefore, must always prevail, more or less, in the management of the affairs of such a company. It is upon this account that joint stock companies for foreign trade have seldom been able to maintain the competition against private adventurers. They have, accordingly, very seldom succeeded without an exclusive privilege, and frequently have not succeeded with one." As Anderson and Tollison (1982) argue, much of Smith's negativity stemmed from empirical observations of the functioning of joint stock companies of his time.

In Smith's time, one of the oldest and largest joint stock companies was the East India Company, which was founded in 1600 for an original period of 15 years, but its desire to extend its monopoly impacted British politics for two centuries. When the British Parliament sought to introduce competition for the East India Company, by giving a charter to one other competitor, some East India Company stockholders simply bought enough shares of the one rival and forced it into a merger, thereby regaining the monopoly position. To seal the deal and prevent future competitive challenges, the East India Company extended a £3.2 million loan to the British Treasury, which, in exchange, again granted the monopoly of trade, allegedly only for a few years. But repeatedly, when the monopoly expired, the East India Company would lobby and pay bribes so that it would be extended—until 1813 for most goods and until 1833 for tea. That a 15-year monopoly right lasted 233 years is a harsh reminder of how dangerous the commingling of economic and political power can be.

Yet the typical high prices and limited output of monopolists was the least of the problems created by the East India Company. The worst aspects were experienced

by people of India and China. By 1764, the East India Company had become the de facto ruler of Bengal, where it established a monopoly in grain trading and prohibited local traders and dealers from "hoarding" rice (which to modern economists looked like a reasonable practice of keeping reserves as insurance against from crop fluctuations). A year after drought struck in 1769, the East India Company raised its already heavy tax on the land. In the aftermath, one out of three Bengalis—more than 10 million people—died of starvation. Another claim to shame for the East India company involved opium. Having lost its monopoly of trade with India (except for tea) in 1813, the East India Company aggressively promoted its export of Bengali opium to China. To defend the right of the East India Company to sell opium to China, the British Empire would wage the two "opium wars."

The Gilded Age

In the heyday of the East India Company, incorporation was a privilege granted only to a few parties by the government. But over time, incorporation became a right of citizens, subject to only a basic registration procedure. This transformation dramatically increased entry in the corporate sector, boosting the degree of competition. But by the late 1800s, another phenomenon contributed to ensuring corporations' market power: the rise in economies of scale, during what Chandler (1977) labels the Second Industrial Revolution. It began with railways, then followed with oil refineries, steel, and chemical production. Great technical achievements were brought about by aggressive entrepreneurs, whom Chernov (1998) calls "titans" and Josephson (1934) calls "robber barons."

In fact, they were both titans and robber barons. In a sports competition, the more disproportionate the reward for the winner vis-à-vis second place, the larger the incentive to take performance-enhancing drugs. Similarly, the more an economy becomes winner-take-all, the bigger the incentives to corrupt the political system to gain a small, but often decisive, advantage. As a result, Chernov's (1998) industrial titans were at the same time the greatest corruptors. In the words of California railway baron Collis Huntington (as quoted in Josephson 1934): "If you have to pay money [to a politician] to have the right thing done, it is only just and fair to do it. ... If a [politician] has the power to do great evil and won't do right unless he is bribed to do it, I think ... it is a man's duty to go up and bribe." Not surprisingly, legal campaign spending reached a peak (in GDP-adjusted terms) at the end of the 19th century (as noted in this journal by Ansolabehere, De Figueiredo, and Snyder 2003). In a cartoon by Joseph Keppler that appeared in *Puck* in 1889, the Senate was labeled "of the Monopolists by the Monopolists and for the Monopolists!"

In a winner-take-all economy, entrepreneurs lobby and corrupt, not only to seize a crucial first-mover advantage, but also to preserve their power over time. They fear political expropriation, which can stem from a populist revolt against the monopolist's abuses or from the rent-seeking of other politically influential parties. This expropriation is made easier when market power does not arise from a technological lead or a skills gap, but from a first-mover advantage or the luck of being

a focal point (such as control over a certain trading venue), because in these cases the transfer implies relatively small deadweight losses.

The overwhelming political power of business was first tamed during the Progressive Era and later by the New Deal. The passage of the Tillman Act in 1907 (which prohibited corporations from making direct contributions to federal candidates) and the Clayton Act in 1914 (which made antitrust enforcement easier) started to limit corporate influence. The New Deal legislation went further, forcing a break up of some of the strongholds of corporate power: investment and commercial banks with the 1933 Glass Steagall Act and corporate pyramids with the 1935 Public Utility Holding Company Act. The coup de grâce, at least for that time period, was provided by the aggressive antitrust enforcement in the later part of Franklin D. Roosevelt's administration, when Thurman Arnold was appointed head of the Antitrust Division of the US Justice Department (Waller 2004).

The Power of Competition and Takeovers

As a result of the political reforms during the first part of the 20th century, the United States entered the second part of the century with a less-concentrated economy. It is not surprising, thus, that economists of that time started to emphasize the limits to firms' power imposed by product market competition and the market for corporate control. For example, Stigler (1958) argued that competitive pressures would determine the scale of firms observed in the marketplace: neither too large nor too small. A competitive selection process in markets for inputs and outputs also eliminates much (if not all) managerial discretion. Thus, in a perfectly competitive industry, as Alchian and Demsetz (1972) wrote: "The firm has no power of fiat, no authority, no disciplinary action any different in the slightest degree from ordinary market contracting between two people."

Even in the absence of competition in product markets, managerial discretion can be constrained by the pressure of the corporate control market. As Manne (1965) first points out, a publicly traded firm that is being run inefficiently represents an arbitrage opportunity. A raider can buy the firm, fix the inefficiency, resell the firm or continue to operate it, and make money.

In these years, neoclassical economics was very successful in moving attention away from the "power" dimension of firms towards the more benign technological aspect. For example, in his prominent microeconomics textbook, Varian (1992) defined firms as "combinations of inputs and outputs that are technologically feasible," and assumes that "a firm acts so as to maximize its profits." When this objective is not assumed, it is derived as a necessary implication of the threat of takeovers (Manne 1965) and intense product market competition (Stigler 1958). Thus, neoclassical economics argued, in a world with perfect competition and no transactions costs, firms are nothing more than isoquant maps.

However, it turns out that even in a perfectly competitive environment, corporations are powerless only if there is perfect contractibility. To understand this idea, we need to depart from the standardized world of neoclassical economics.

The Incomplete Contract Paradigm

The neoclassical framework only describes well one set of transactions, called "standardized" transactions by Williamson (1985), which involve many producers of similar quality products and many potential customers. Many common transactions, however, do not fit this mold.

For instance, consider the purchase of a customized machine. The buyer must contact a manufacturer and agree upon the specifications and the final price. More importantly, signing the agreement does not represent the end of the relationship between the buyer and the seller. Producing the machine requires time. During this time, events can occur that alter the cost of producing the machine as well as the buyer's willingness to pay for it. Before the agreement was signed, the market for manufacturers may have been competitive. Once production has begun, though, the buyer and the seller are trapped in a bilateral monopoly. The customized machine probably has a higher value to the buyer than to the market. On the other hand, the contracted manufacturer probably has the lowest cost to finish the machine. The difference between what the two parties generate together and what they can obtain in the marketplace represents a quasi-rent, which needs to be divided. Of course, the initial contract plays a main role in dividing this quasi-rent. But most contracts are incomplete, in the sense that they will not fully specify the division of surplus in every possible contingency (this might be too costly to do or even outright impossible because the contingency was unanticipated). This creates an interesting distinction between decisions made at the time when the two parties entered a relationship and irreversible investments were sunk, and decisions made later in the process when the quasi-rents are divided.

The incompleteness of the contract creates room for bargaining. The outcome of the bargaining will be affected by several factors besides the initial contract. First, it will depend to some extent on which party has ownership of the machine while it is being produced. Second, it will depend on the availability of alternatives: How costly is it for the buyer to delay receiving the new machine. How costly is it for the seller of the machine to delay the receipt of the final payment? How much more costly is it to have the job finished by another manufacturer? Finally, the institutional environment plays a major role in shaping the bargaining outcome: How effective and rapid is law enforcement? What are the professional norms? How quickly and reliably does information about the manufacturer's performance travel across potential clients? All these factors determine the allocation of authority or power. In this setting, given that not all contingencies can be specified, what is often specified instead is who has the right to make decisions when unspecified contingencies arise, which in turn will influence strategic bargaining over the surplus. In this context, Grossman and Hart's (1986) "residual right of control" is both meaningful and valuable.

Extending the Incomplete Contract Paradigm

The incomplete contract literature started by Grossman and Hart (1986) explains how firms' power stems from their market power (Rajan and Zingales

1998). While it focuses only on the market power arising from past investments, this link also holds when the source of market power is economies of scale, network externalities, or government-granted licenses (Rajan and Zingales 2001).

Furthermore, emphasizing the incomplete nature of contracts and rules, the theory of incomplete contracts creates scope for lobbying, rent seeking, and power grabbing. The traditional contributions focus on the under- or overinvestments in firm-specific human capital, but the framework can easily be extended to the political arena. If rents are not perfectly allocated in advance by contracts and rules, there is ample space for economic actors to exert pressure on the regulatory, judiciary, and political system to grab a larger share of these rents.

As far as I know, the interaction between these dimensions has thus far gone largely unstudied. In a world where cash bribes are illegal and relatively rare, firms need other means to lobby and pressure the political and regulatory world. One common mechanism is, for example, the (implicit) promise of future career opportunities. The credibility (and thus the effectiveness) of such promises strongly depends upon the current and future economic power of a firm. At the peak of the financial crisis, Citigroup offers were not very credible, because there were serious doubts that Citigroup would survive. By contrast, JP Morgan Chase chief executive officer Jamie Dimon was seen as a reliable player, because of the staying power of his bank. Thus, even without mobilizing its finances, the more economically powerful a firm is, the more politically powerful it can be.

If the *ability* to influence the political power increases with economic power, so does the *need* to do so, because the greater the market power a firm has, the greater the fear of expropriation by the political power. Hence, the risk of what I will call the "Medici vicious circle."

The Medici Vicious Circle

A competitive advantage often starts as temporary. The video rental chain Blockbuster was founded on the idea that videos had become a mainstream product, which no longer needed to be rented in shady stores full of compromising material, and could instead be rented in a family-friendly setting with bright lights and a vivid store logo. This simple (and replicable) idea was quickly transformed into a network of stores across cities. Once the network of local stores was in place, Blockbuster had a huge barrier to entry vis-à-vis any competitor, but a barrier that was eventually overcome by the technology of accessing and renting movies over the internet.

Most firms are actively engaged in protecting their source of competitive advantage through a mixture of innovation, lobbying, or both. As long as most of the effort is along the first dimension, there is little to be worried about. The fear of being overtaken pushes firms to innovate (Aghion, Akcigit, and Howitt 2013). What is more problematic is when a lot of effort is put into lobbying.

In other words, the problem here is not temporary market power. The expectation of some temporary market power based on innovation is the driver of much innovation and progress. The fear is of a "Medici vicious circle," in which money is

used to gain political power and political power is then used to make more money.[1] This vicious circle needs to be broken. In the case of medieval Italy, this cycle turned Florence from one of the most industrialized and powerful cities in Europe to a marginal province of a foreign empire. At least the Medici period left some examples of great artistic beauty in Florence. I am not sure that market capitalism of the 21st century will be able to do the same.

The Increasing Market Power of US Firms

In a perfectly competitive world, the economic power of firms stems only from the past specific investments. The potential magnitude of this economic power is limited and does not benefit much from the support of political power. In this Economics 101 world, lobbying is an activity limited to firms that are trying either to escape from the jaws of regulation or to attract government contracts. In this setting, the neoclassical description of the firm as having "no power of fiat, no authority" is a reasonable approximation of reality.

One can argue whether such a close-to-competitive economy ever existed, but one cannot argue that this is the world we live in today, even in the United States, which historically has done fairly well relative to many other countries along this dimension. In the last two decades, more than 75 percent of US industries experienced an increase in concentration levels, with the Herfindahl index increasing by more than 50 percent on average. During this time, the size of the average publicly listed company in the United States tripled in market capitalization: from $1.2 billion to $3.7 billion in 2016 dollars (Grullon, Larkin, and Michaely 2017; see also the discussion by Kahle and Stulz in this symposium).

This phenomenon is the result of two trends: On the one hand, the reduction in the rate of birth of new firms, which went from 14 percent of existing firms in the late 1980s to less than 10 percent in 2014 (Haltiwanger 2016); on the other hand, a very high level of merger activity, which for many years in the last two decades exceeded $2 trillion in value per year (Institute for Mergers, Acquisitions & Alliances, at https://imaa-institute.org/mergers-and-acquisitions-statistics/).

Impact on Margins

Higher concentration does not mean necessarily higher market power, yet there is increasing evidence that market power has increased. First of all, these mergers do seem to improve productivity, but only to raise mark-ups from 15 percent to over 50 percent of the average markup (Blonigen and Pierce 2016). The market power enjoyed by larger firms is also reflected in the increasing difficulty that smaller firms

[1] Several recent secondary sources claim that the Medici family motto was: "Money to get power. Power to protect money." However, none of these sources offers a primary attribution. For example, this claim appears in the Santi (2003) book of quotations, in the 2005 movie "The American Ruling Class" (as discussed in Walton 2011), and in Gross (1980).

have in competing in the marketplace: in 1980, only 20 percent of small publicly traded firms had negative earnings per share; in 2010, 60 percent did (Gao, Ritter, and Zhu 2013).

The most convincing evidence on this theme is provided by Barkai (2016), who finds that the decrease in labor share of value added is not due to an increase in the capital share (that is, the cost of capital times amount of capital divided by value added), but by an increase in the profits share (the residuals), which goes from 2 percent of GDP in 1984 to 6 percent in 2014. This is not just a relabeling. By separating the return to capital and profits, we can discern when profits come from (nonreplicable) barriers to entry and competition rather than from capital accumulation. Distinguishing between capital and profit share allows Barkai also to gain some insights on the cause of the decline in the labor share. If markups (the difference between the cost of a good and its selling price) are fixed, any change in relative prices or in technology that causes a decline in labor share must cause an equal increase in the capital share. If both labor and capital share dropped, then there must be a change in markups—that is, the pricing power of firms to charge more than their cost. In support of this "market power" hypothesis, Barkai finds that sectors that have experienced a higher increase in concentration between 1997 and 2012 also experienced a higher decline in labor share of output and thus (presumably) a higher increase in the share of profits.

Possible Explanations

A first popular explanation for these trends is the emergence and diffusion of network externalities: that is, situations in which an increase in usage leads to a direct increase in value for other users. These externalities have been present at least since the telephone, but they have become much more widespread with the diffusion of the internet and of social media.

A second explanation is the increased role of winner-take-all industries, driven by the proliferation of information-intensive goods that have high fixed and low-marginal costs (Zingales 2012; Autor, Katz, Patterson, and Van Reenen 2017). A related explanation has to do with information complementarities. The value of the data derived from Facebook and Instagram combined is likely to be higher than the sum of the value of the data derived from Facebook and Instagram separately, since the data can be combined and compared. Thus, Facebook is likely to be the higher-value user of Instagram data, even ignoring any potential market power effect. If you add market power effects, the momentum toward concentration might be irresistible.

A final explanation is reduced antitrust enforcement. Section 2 of the Sherman Act makes it unlawful to "monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce." During the period 1970–1999, the Department of Justice and the Federal Trade Commission (FTC) together brought an average of 15.7 cases under Section 2. Between 2000 and 2014, they brought only 2.8 cases a year (Grullon,

Larkin, and Michaely 2017). These explanations are not mutually exclusive; in fact, they are mutually reinforcing.

Political Power of Firms

There are many misconceptions about the nature and the importance of political power of firms. If politics is identified along partisan lines, corporations are not very relevant, nor do they want to be. Rich individuals, like casino magnate Shelden Adelson, play a big role in funding political campaigns; corporations do not. As Ota (1998) reported: "'Mickey Mouse is not a Republican or a Democrat,' said Joe Shapiro, who oversaw Disney's Washington lobbying office in the early 1990s. 'If you take a strong position either way, you are looking at offending roughly half of the people.'" The secondary impact of corporations in determining which party prevails explains how Donald Trump could be elected to the presidency in 2016 despite not having the endorsement (and the money) from political action committees at any of the top 100 US corporations.

Corporations need some friends in Congress (and in the executive branch) on specific issues, and they generally succeed in having them, regardless of the political affiliation. Consider Citigroup's effort to change the Glass–Steagall Act, which severed the economic ties between investment banking and commercial banking. In 1998, Citigroup acquired Travelers (an insurance company), even though the law prohibited banks from merging with insurance companies. At the time of the merger, Travelers' CEO, Sandy Weill (as reported in Martin 1998), explained why the companies were moving forward in spite of an apparent conflict with the law: "[W]e have had enough discussions [with the Fed and the Treasury] to believe this will not be a problem." The head of the US Treasury then was Richard Rubin, who worked very hard to convince his fellow Democrats to change the law. Rubin left the Treasury in July 1999, the day after the House of Representatives passed its version of the bill by a bipartisan vote of 343 to 86. Three months later, on October 18, 1999, Rubin was hired at Citigroup at a salary of $15 million a year, without any operating responsibility.

Even when it comes to lobbying, the actual amount spent by large US corporations is very small, at least as a fraction of their sales. For example, in 2014 Google (now Alphabet) had $80 billion in revenues and spent $16 million in lobbying (see the Lobbying Database at OpenSecrets.org, https://www.opensecrets.org/lobby). To the extent that US corporations are exercising political influence, it seems that they are choosing less-visible but perhaps more effective ways. In fact, since Gordon Tullock's (1972) famous article, it has been a puzzle in political science why there is so little money in politics (as discussed in this journal by Ansolabehere, de Figueiredo, and Snyder 2003).

One possible explanation is that corporations do not need as much to prevail politically because the opposition they face (which might be broadly understood as the interest of the general public) is very disorganized and they can prevail with very

little effort (Zingales 2012, chap. 5). If money is used only in the marginal cases, one can observe very little correlation between donations and success (Ansolabehere, de Figueiredo, and Snyder 2003). However, it certainly seems in specific cases that big corporations have a high success rate in getting their wishes to come true; for example, see Pierson's (2015) discussion of the health care reform legislation.

Another explanation is that actual donation amounts and lobbying are so small because big corporations are so good at achieving their goals without the need of cash transfers. Nobody would try to measure the influence of the Mafia with the size of the bribes they pay. In fact, the power of a boss, like Vito Corleone in Mario Puzo's 1969 book *The Godfather*, does not rest on his ability to pay, but on his power to make offers to people that they cannot refuse. Of course, the Mafia relies on not-so-veiled threats of violence, while corporate interests do not. Yet, the successful Mafia boss is able to minimize violence: it is an out-of-equilibrium threat, rarely carried through. Corporate interest can use a threat of ostracism from the business world at the end of a public official's mandate. That such ostracism is rarely observed is consistent with the belief that it is a highly effective threat.

In other words, to detect the power of corporations we need to look at output, not inputs. Is it a coincidence that the common term of copyright is extended every time the copyright of the Walt Disney Company on Mickey Mouse is close to expiration (Lessig 2001)? This case is so outrageous not because it is so unique, but because there is no ideological cover for it: extending retroactively copyright to long-dead authors is not likely to stimulate production of new works!

Similarly, we can ask why the antitrust case of the Federal Trade Commission against Google was dropped in the United States, while parallel efforts were not dropped in Europe. A leaked FTC staff report (available via the *Wall Street Journal* website at https://graphics.wsj.com/google-ftc-report/img/ftc-ocr-watermark.pdf) concluded that Google had unlawfully maintained its monopoly over general search and search advertising by "scraping content from rival vertical websites," "by entering into exclusive and highly restrictive agreements with web publishers that prevent publishers from displaying competing search results or search advertisements," and "by maintaining contractual restrictions that inhibit the cross-platform management of advertising campaigns." Nonetheless, the FTC unanimously decided to drop the case. One wonders if the frequent visits paid by Google employees to the White House played a role: between Obama's first inauguration and the end of October 2015, employees of Google and associated entities visited the White House 427 times, including 21 small, intimate meetings with President Obama (as reported by the Google Transparency Project at http://googletransparencyproject.org/articles/googles-white-house-meetings).

From Brown and Huang (2017), we learned that the share price of companies whose executives visited the White House from 2009–2015 increased an extra 1 percent in the following two months. It might not seem very much, until you discover that during Obama's presidency, the chairman and chief executive officer of Honeywell international visited the White House 30 times, while the head of General Electric visited 22 times.

If companies do not succeed in preventing unfavorable legislation in Congress, they can stop it by suing the regulators who try to implement it. The Dodd–Frank Wall Street Reform and Consumer Protection Act of 2010 required that the US Securities and Exchange Commission repeal its rules that prevent institutional investors from nominating their own representatives to corporate boards. In fact, the requirement was very timid, posing so many restrictions in terms of quantity and length of ownership as to leave the bar to institutional investors effectively in place. Still, the Business Roundtable sued the SEC to block the rule. The case was argued by Eugene Scalia, the son of then-US Supreme Court Justice Antonin Scalia, and was won in the US Court of Appeals, DC Circuit, on a technicality—the failure of the SEC to conduct a cost–benefit analysis ahead of time. This small setback turned into a major defeat for shareholders when the SEC, rather than performing such an analysis and re-proposing the rule, chose to withdraw. At a conference in December 2011, I asked then-SEC Chairwoman Mary Schapiro when her agency was planning to reintroduce the rule. I even offered to do the cost–benefit analysis for free. But she confessed that the SEC had many other items on its agenda, and had placed the issue on the back burner, which seems to me a polite way of saying that the SEC had surrendered under pressure.

If all else fails, large companies can succeed in avoiding regulation by lobbying the regulator directly, so as to avoid enforcement. Lambert (2015) finds that regulators are 44.7 percent less likely to initiate enforcement actions against lobbying banks.

Lobbying is not the only way companies have to avoid enforcement: they can do so by hiding crucial information. As described in Shapira and Zingales (2017), DuPont was able to delay by more than 30 years any liability for contaminating the water supply near its West Virginia factory, by hiding information and protecting itself behind the trade secret law.

Why the Problem Is Getting Worse

All the actions described above require not only money, but also power of fiat and disciplinary action, which differ from ordinary market contracting between two economic actors. Thus, in a fragmented and competitive economy, firms find it difficult to exert this power. In contrast, firms that achieve some market power can lobby (in the broader sense of the term) in a way that ordinary market participants cannot. Their market power gives them a comparative advantage at the influence game: the greater their market power, the more effective they are at obtaining what they want from the political system. Moreover, the more effective they are at obtaining what they want from the political system, the greater their market power will be, because they can block competitors and entrench themselves. Hence, the risk of a Medici vicious circle.

In the last three decades in the United States, the power of corporations to shape the rules of the game has become stronger for three main reasons. First, the size and market share of companies has increased, which reduces the competition across conflicting interests in the same sector and makes corporations more

powerful vis-à-vis consumers' interest. Second, the size and complexity of regulation has increased, which makes it easier for vested interests to tilt the playing field to their advantage. Finally there has been a demise of the antibusiness ideology that previously prevailed among Democrats, and this has reduced the costs of being perceived as too friendly to the interests of big business for both parties.

An example of this increased power of corporations is found in the legislative history of the 2005 Bankruptcy Abuse Prevention and Consumer Protection Act described in Zingales (2012, chapter 4). In the words of one legal scholar: "Never before in our history has such a well-organized, well-orchestrated, and well-financed campaign been run to change the balance of power between creditors and debtors." (Tabb 2007).

Towards a Political Theory of the Firm

It is not at all my intention to conclude that business should have no voice in the political process. After all, other powerful special interests, such as unions, play a role in politics. Without some corporate voice, the outcome of political decisions could become too tilted in other directions. Even more importantly, the power of the state over its citizens might become excessive without a strong constituency that defends property rights.

The ideal state of affairs is a "goldilocks" balance between the power of the state and the power of firms. If the state is too weak to enforce property rights, then firms will either resort to enforcing these rights by themselves (through private violence) or collapse. If a state is too strong, rather than enforcing property rights it will be tempted to expropriate from firms. When firms are too weak vis-à-vis the state, they risk being expropriated, if not formally (with a transfer of property rights to the government), then substantially (when the state demands a large portion of the returns to any investment). But when firms are too strong vis-à-vis the state, they may shape the definition of property rights and its enforcement in their own interest and not in the interest of the public at large, as in the Mickey Mouse Copyright Act example. The feasibility of a "goldilocks" equilibrium depends upon a mixture of institutional and economic characteristics.

By this metric, the United States does relatively well in international and historical comparisons. This fortune, however, should not be taken for granted. The Second Industrial Revolution, at the end of the 19th and start of the 20th century, upset the "goldilocks" equilibrium, which was only restored with great effort over four decades of reforms. The Third Industrial Revolution now underway is having similar effects. To understand this phenomenon in an historical and international context, it is useful to review the different types of equilibria present around the world.

Institutional Characteristics

In most autocratic regimes, the main source of power is the control of the armed forces and police—that is, a monopoly over the legal use of force. In such

countries, de facto control of the armed forces greatly depends upon the personal loyalty of the rank-and-file soldiers to some commanders and the future rents a leader can credibly promise. By contrast, in democratic regimes the source of political power is a broad social consensus, formalized through an election process.

One key mechanism in the formation of this democratic consensus is the world of the media, itself influenced by the political power (through censorship, ownership, subsidies, and leaks) and by the economic power (through advertising, direct ownership, financing, and access to information). Traditionally, media have been considered free if they were not affected by government censorship. Yet, it is equally important that they are (mostly) not affected by corporate censorship, which can be a frequent phenomenon especially in small countries, where the media market is often controlled by few well-connected families (for example, Zingales 2016).

A second key mechanism in the formation of political consensus is the electoral process, shaped both by the electoral law and by the rules for campaign financing. A more proportional system of representation favors new entry and competition, but it also makes it easier for vested interests to capture small parties and turn them into lobbying organizations for special interests. The source of campaign financing is also crucial. When campaign financing comes from the government, the political control is greater; when it comes from private donations, economic power can be greater. A mixture of limitations on private donations, matched to some extent by public financing, is an attempt to find a balance between these alternatives.

In the formation of consensus and in legitimizing the political authority, a third factor is the role of ideology. In some countries, political legitimization is linked to a formal election process; in other countries, governments formed in different ways are nonetheless regarded as legitimate. Ideology is also based on perceptions of the relative benefits of being dominated by economic interests.

Finally, a crucial role is played by the prosecutorial and judiciary powers. These differ in their degree of independence from the political and the economic powers and in their prevalent ideology. For example, Epstein, Landes, and Posner (2013) document the big increase in pro-business decisions in the US Supreme Court between 1946 and 2011.

Economic Characteristics

Given the legal and social restrictions on explicit bribes in most countries, a company's ability to obtain what it wants from the political system is highly dependent upon: 1) its ability to make credible long-term promises (for example, future employment opportunities for politicians and regulators), which is highly dependent upon a company's long-term survival probability; 2) the grip a company has on the market for specific human capital (for example, how many potential employers of nuclear engineers there are); 3) a company's ability to wrap its self-interest in a bigger, noble, idea (for example, Fannie Mae and the goal that every American should be able to borrow to purchase a house); 4) the control that a company has through its image in society by way of employment, data ownership, media ownership, advertising, research funding, and other methods.

In economic terms, a firm's size and the level of concentration within a market affect positively all the crucial factors that determine a firm's ability to influence the political system. What matters here it is not just product market concentration, but in general all concentration of economic power. The main employer in a town or jurisdiction is very politically influential, even if the firm sells in a competitive market outside that town.

A Taxonomy

If we focus on the balance between economic and political power, we can identify some prototypical regimes. At the one extreme, there are traditional communist dictatorships, like the old Soviet Union, as well as North Korea and Cuba. In a communist dictatorship, political power has captured all important sources of economic power. At the other extreme, there is the most extreme form of plutocratic regimes, like the East India Company protectorate of India or the King Leopold II ownership of Congo. In such cases, the economic power has captured the functions of the political power. The "banana republics" of the early 20th century (the term is used to describe how large US firms like United Fruit Company created a near-monopoly supply of bananas from countries in Latin America and the Caribbean) were a modified version of these pure plutocracies. At least formally, the banana republic countries had an alternative source of political power, while the East India Company system in India before 1858 and the Congo Free State before 1908 did not.

Moving towards the center we find two types of regimes that, while very different in their nature, tend to be similar in their outcomes. On the one hand, we find the political patronage regimes of Suharto in Indonesia, Goodluck Johnathan in Nigeria, and many heads of government in Africa today. In these regimes, political power grants economic power through methods like concessions of either mineral extraction rights or monopoly (or quasi-monopoly) rights to operate certain businesses. A special mention is due to Egypt, where the army has transformed itself into a conglomerate, running all sort of commercial enterprises on the side.

On the other hand, we have the "vertical politically integrated" regimes (Haber, Razo, and Maurer 2003), where rich businessmen control the political system, sometimes directly (as was the case in Thailand under Thaksin Shinawatra and Italy under Silvio Berlusconi) or sometimes indirectly (as the Russian oligarchs under Vladimir Putin). These regimes differ in the degree of concentration in the main source of power. Suharto in Indonesia or Robert Mugabe in Zimbabwe had close to a monopoly grip on political power, while Goodluck Johnathan in Nigeria did not. In the same way, British Petroleum before its Middle Eastern operations were nationalized had close to a monopoly on the sources of economic power in Iran, while oligarchs in Russia and Berlusconi in Italy did not. While the original source of power is very different, political patronages and vertical politically integrated regimes are very similar in the way they use the political power to protect and enhance business. In fact, countries often oscillate between these regimes: for example, Russia moved from a vertical politically integrated regime under Boris Yeltsin, to a political patronage under Putin.

While the perfect "goldilocks" balance is an unattainable ideal, given that ongoing events will expose the tradeoffs in any given approach, the countries closest to this ideal are probably the Scandinavian countries today and the United States in the second part of the twentieth century. Crucial to the success of a goldilocks balance is a strong administrative state, which operates according to the principal of impartiality (Rothstein 2011) and a competitive private sector economy.

In Scandinavian countries, the competitiveness of the sector is ensured by the small size of these countries, which forces them to be open and subjected to international competition. The quality of government is ensured by a long tradition of benign and enlightened monarchies that have evolved smoothly into democracies, along with ethnic homogeneity that favors an identification of the citizens with the state.

Historically, competition in the United States was ensured by the very large size of the country relative to the size of the then-existing companies and the ability of their managers to travel and congregate, which made it more difficult for a small group of producers to "own" the government. During the Cold War period after World War II, the efficiency of the government was required by the threat of military conflict. Both these aspects have diminished now, increasing the risk that the United States becomes a vertical politically integrated regime with greater similarity to some countries of Latin America.

Conclusion

In a famous speech in 1911, Nicholas Murray Butler, President of Columbia University, considered the practical advances made by large corporations in the late 19th and early 20th century and stated:

> I weigh my words, when I say that in my judgment the limited liability corporation is the greatest single discovery of modern times, whether you judge it by its social, by its ethical, by its industrial or, in the long run,—after we understand it and know how to use it,—by its political, effects. Even steam and electricity are far less important than the limited liability corporation, and they would be reduced to comparative impotence without it.

Butler was right, but incomplete. This discovery of the modern corporate form—like all discoveries—can be used to both to foster progress or to oppress. The size of many corporations exceeds the modern state. As such, they run the risk of transforming small- and even medium-sized states into modern versions of banana republics, while posing economic and political risks even for the large high-income economies.

To fight these risks, several political tools might be put into use: increases in transparency of corporate activities; improvements in corporate democracy; better rules against revolving doors and more attention to the risk of capture of scientists

and economists by corporate interests; more aggressive use of the antitrust authority; and attention to the functioning and the independence of the media market. Yet the single most important remedy may be broader public awareness. Without an awareness of this risk of deterioration of the corporate form, and a sense of how to strike the appropriate balance between corporations and governments, there is little hope for any remedy.

■ *I thank Mark Gertler, Gordon Hanson, Enrico Moretti, and Timothy Taylor for very useful comments and Steve Haber for educating me on the multiple ways in which Google lobbies. Financial support from the Stigler Center and from the George Rinder Research fund at the University of Chicago is gratefully acknowledged.*

References

Aghion, Philippe, Ufuk Akcigit, and Peter Howitt. 2013. "What Do We Learn from Schumpeterian Growth Theory?" PIER Working Paper 13–026.

Alchian, Armen A., and Harold Demsetz. 1972. "Production, Information Costs, and Economic Organization." *American Economic Review* 62(5): 777–95.

Anderson, Gary M., and Robert D. Tollison. 1982. "Adam Smith's Analysis of Joint-Stock Companies." *Journal of Political Economy* 90(6): 1237–56.

Ansolabehere, Stephen, John M. de Figueiredo, and James M. Snyder Jr. 2003. "Why Is There So Little Money in U.S. Politics?" *Journal of Economic Perspectives* 17(1): 105–30.

Autor, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2017. "Concentrating on the Fall of the Labor Share." *American Economic Review* 107(5): 180–85.

Barkai, Simcha. 2016. "Declining Labor and Capital Shares." http://home.uchicago.edu/~barkai/doc/BarkaiDecliningLaborCapital.pdf.

Berle, Adolf A., and Gardiner Means. 1932. *Modern Corporation and Private Property.* Piscataway, NJ: Transaction Publishers.

Blonigen, Bruce A., and Justin R. Pierce. 2016. "Evidence for the Effects of Mergers on Market Power and Efficiency." Federal Reserve Board Divisions of Research & Statistics and Monetary Affairs Finance and Economics Discussion Paper 2016–082.

Brown, Jeffrey R., and Jiekun Huang. 2017. "All the President's Friends: Political Access and Firm Value." NBER Working Paper 23356.

Butler, Nicholas Murray. 1911. "Politics and Business: Address of Nicholas Murray Butler Ph.D., D.Litt. LL.D., President of Columbia University." Lecture presented at the 143rd Annual Banquet of the Chamber of Commerce of the State of New York, New York, NY, November 16, 1911.

Chandler, Alfred D. 1977. *The Visible Hand: The Managerial Revolution in American Business.* Harvard University Press.

Chernow, Ron. 1998. *Titan: The Life of John D. Rockefeller, Sr.* New York: Random House.

Committee on Capital Market Regulation. 2014. *Annual Shareholder Meetings and the Conundrum of "Unelected" Directors.* Cambridge, MA: Committee on Capital Markets Regulation.

Epstein, Lee, William M. Landes, and Richard A. Posner. 2013. "How Business Fares in the Supreme Court." *Minnesota Law Review* 97(1): 1431–72.

Gao, Xiaohui, Jay R. Ritter, and Zhongyan Zhu. 2013. "Where Have All the IPOs Gone?" *Journal of Financial and Quantitative Analysis* 48(6): 1663–92.

Global Justice Now. 2016. "10 Biggest Corporations Make More Money than Most Countries in the World Combined." September 12.

http://www.globaljustice.org.uk/news/2016/sep/12/10-biggest-corporations-make-more-money-most-countries-world-combined.

Google Transparency Project. No date. "Google's White House Meetings." http://googletransparencyproject.org/articles/googles-white-house-meetings.

Gross, Bertram. 1980. *Friendly Fascism: The New Face of Power in America*. New York: South End Press.

Grossman, Sanford J., and Oliver D. Hart. 1986. "The Costs and the Benefits of Ownership: A Theory of Vertical and Lateral Integration." *Journal of Political Economy* 94(4): 691–719.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely. 2017. "Are U.S. Industries Becoming More Concentrated?" Available at SSRN: https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2612047.

Haber, Stephen, Armando Razo, and Noel Maurer. 2003. *The Politics of Property Rights: Political Instability, Credible Commitments, and Economic Growth in Mexico, 1876–1929*. Cambridge University Press.

Haltiwanger, John C. 2016. "Firm Dynamics and Productivity: TFPQ, TFPR, and Demand-Side Factors." *Economia Journal of the Latin American and Caribbean Economic Association* 17(1): 3–26.

Jensen, Michael C., and William H. Meckling. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *Journal of Financial Economics* 3(4): 305–60.

Josephson, Matthew. 1934. *The Robber Barons*. New York: Houghton Mifflin Harcourt.

Keppler, Joseph. 1889. *The Bosses of the Senate*, a cartoon in *Puck,* January 23. https://commons.wikimedia.org/wiki/File:The_Bosses_of_the_Senate_by_Joseph_Keppler.jpg.

Lambert, Thomas. 2015. "Lobbying on Regulatory Enforcement Actions: Evidence from U.S. Commercial and Savings Banks." Available at SSRN: https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2517235.

Lessig, Lawrence. 2001. "Copyright's First Amendment." *UCLA Law Review* 48: 1057–65.

Manne, H. G. 1965. "Mergers and the Market for Corporate Control." *Journal of Political Economy* 73: 110–20.

Martin, Mitchell. 1998. "Citicorp and Travelers Plan to Merge in Record $70 Billion Deal: A New No. 1: Financial Giants Unite." *New York Times*, April 7. http://www.nytimes.com/1998/04/07/news/citicorp-and-travelers-plan-to-merge-in-record-70-billion-deal-a-new-no.html.

OpenSecrets.org. No date. Lobbying Database. http://www.opensecrets.org/lobby/.

Ota, Alan K. 1998. "Disney in Washington: The Mouse That Roars." *Congressional Quarterly This Week (CNN)*, August 10. http://edition.cnn.com/ALLPOLITICS/1998/08/10/cq/disney.html.

Pierson, Paul. 2015. "Goodbye to Pluralism? Studying Power in Contemporary American Politics." Paper presented at the Wildavsky Forum for Public Policy, Goldman School of Public Policy, Berkeley, CA, April 9, 2015.

Puzo, Mario. 1969. *The Godfather*. G. P. Putnam's Sons.

Rajan, Raghuram G., and Luigi Zingales. 1998. "Power in a Theory of the Firm." *Quarterly Journal of Economics* 113(2): 387–432.

Rajan, Raghuram G., and Luigi Zingales. 2001. "The Firm as a Dedicated Hierarchy: A Theory of the Origins and Growth of Firms." *Quarterly Journal of Economics* 116(3): 805–51.

Rothstein, Bo. 2011. *The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective*. Chicago: University of Chicago Press.

Stigler, G. J. 1958. "The Economies of Scale." *Journal of Law and Economics* 1: 54–71.

Santi, Antonio. 2003. *The Book of Italian Wisdom*. New York: Citadel Press.

Shapira, R., and Luigi Zingales. 2017. "Is Pollution Value Maximizing? The DuPont Case." Unpublished paper.

Smith, Adam. 1776 [1904]. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Edited by Edwin Cannan. London: Methuen & Co., Ltd.

Tabb, Charles Jordan. 2007. "The Top Twenty Issues in the History of Consumer Bankruptcy." *University of Illinois Law Review* 2007(1): 9–30.

Tullock, Gordon. 1972. "The Purchase of Politicians." *Western Economic Journal* 10: 354–55.

Varian, Hal R. 1992. *Microeconomic Analysis*, 3rd edition. New York: W. W. Norton & Company.

Waller, Spencer Weber. 2004. "The Antitrust Legacy of Thurman Arnold." *St. John's Law Review* 78(3): 569–614.

Walton, Beatrice. 2011. "The American Ruling Class." *Harvard Political Review*, September 6. http://harvardpolitics.com/books-arts/the-american-ruling-class/.

Williamson, Oliver E. 1985. *The Economic Institutions of Capitalism*. New York: Free Press.

Zingales, Luigi. 2012. *A Capitalism for the People: Recapturing the Lost Genius of American Prosperity*. New York: Basic Books.

Zingales, Luigi. 2016. "Are Newspapers Captured by Banks? Evidence From Italy." *ProMarket,* May 12. https://promarket.org/are-newspapers-captured-by-banks/.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 131–150

A Skeptical View of Financialized Corporate Governance

Anat R. Admati

The vast bulk of economic activity today involves business corporations. Corporations are abstract legal entities that combine legal rights and obligations with a significant degree of flexibility. The legal separation between corporations and their stakeholders, including shareholders, has been important to the success of the corporate form in organizing long-term, large-scale production, while limited liability and the tradability of shares help corporations acquire funds from a broad set of investors.

However, this legal separation exacerbates conflicts of interest between those who control corporations and others, including shareholders, creditors, employees, suppliers, customers, public authorities, and the general public. In large corporations, stakeholders vary enormously in the information and degree of control they have on corporate actions. Contracts and markets do not generally create efficient outcomes if markets are not competitive, contracts are incomplete or costly to enforce, or if corporate actions create negative externalities for those with little information or control. Laws and regulations can help alleviate these frictions, but their design and enforcement are also costly and subject to information and control frictions.

In recent decades, much emphasis has been placed on aligning the interests of managers and shareholders. Managerial compensation typically relies on financial yardsticks such as profits, stock prices, and return on equity to achieve

■ *Anat R. Admati is the George G. C. Parker Professor of Finance and Economics, Graduate School of Business, Stanford University, Stanford, California. Her email address is admati@stanford.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.131 doi=10.1257/jep.31.3.131

such alignment. This development has been part of a broader trend referred to as "financialization," whereby the financial sector and financial activities grow in prominence within the economy, and financial markets and measures increasingly guide economic activity.

Financialized governance may not actually work well for most shareholders. Even when financialized governance benefits shareholders, significant tradeoffs and inefficiencies can arise from the conflict between maximizing financialized measures and society's broader interests. For example, financialized governance provides incentives for slanted presentations of accounting data and even in some cases outright accounting fraud. Misconduct, law evasion, or fraud directed at other stakeholders such as customers and governments may benefit shareholders, but they may ultimately have to bear legal expenses, large fines, and loss of reputation. Financialized incentives can also lead to misallocation through "short-termism" or mismanagement of risk, with the upside benefiting those controlling corporations and the downside harming others, including shareholders and the broader economy.

Effective governance requires that those in control are accountable for actions they take. However, those who control and benefit most from corporations' success are often able to avoid accountability. In cases such as corporate fraud or excessive endangerment in which the public is insufficiently aware of the potential conflicts, governments may fail to design and enforce the best rules because of the incentives of individuals within governments and their own lack of accountability.

The important real-world issues around corporate governance do not fit neatly into most common economic frameworks and models. The history of corporate governance includes a parade of scandals and crises that have caused significant harm. Although each episode has its unique elements, after each scandal or crisis, the narratives of most key individuals tend to minimize their own culpability or the possibility that they could have done more to prevent the problem. Common claims from executives, boards of directors, auditors, rating agencies, politicians, and regulators include "we just didn't know," "we couldn't have predicted," or "it was just a few bad apples." A recent report commissioned by the board of directors of Wells Fargo Bank regarding the scandal in which bank employees misled customers and fraudulently opened accounts for years referred to executives and the board as having a "disinclination ... to see the problem as systemic" despite numerous flags and opportunities to act (Independent Directors 2017, p. 6).

Economists, as well, may react to corporate scandals and crises with their own version of "we just didn't know," as their models had ruled out certain possibilities. They may interpret events as benign, arising from exogenous forces out of anybody's control, or try to fit the observations into alternative models. However, many economic models still ignore highly relevant issues of incentives, governance conflicts, enforcement, and accountability. Economists may presume that observed reality is unchangeable or efficient under one set of frictions, while leaving out other frictions and ways to address them through changes in governance practices or policy.

Effective governance of institutions in the private and public sectors should make it much more difficult for individuals in these institutions to get away with claiming that harm was out of their control when in reality they had encouraged or enabled harmful misconduct and, moreover, when they could have and should have taken action to prevent it. Better practices and policy would follow.

Financialization and Shareholder Governance

The last few decades have seen an expansion of financial activity and financial markets driven by a number of factors: increased volatility of exchange rates and interest rates, globalization, changes in financial regulations, and financial innovations such as securitization and derivatives (Davis 2011; Krippner 2011). The expansion of financial activity has offered greater risk-sharing opportunities and enabled innovations, large-scale investments, and economic growth. However, it has also allowed risk to become hidden and magnified in an opaque and complex system that is rife with conflicts of interest (Partnoy 2009; Zingales 2015). Whereas economists usually presume that the size of a sector is efficient if it is determined in markets, recent empirical work argues that "too much finance" may harm growth, create distortions, and contribute to income inequality (Cecchetti and Kharroubi 2015; Cournède and Denk 2015; de Haan and Sturm 2016).

My focus here is on the interaction of financialization and corporate governance. Financialized corporate governance starts with the view, especially dominant in the United States and the United Kingdom, that corporations should focus on benefiting shareholders (Hansmann and Kraakman 2001). The economics and finance literatures have focused almost exclusively on potential conflicts of interest between shareholders and managers (Bebchuk and Weisbach 2010). In recent decades, the main approach to resolve that conflict has been to incentivize a maximization of "shareholder value" by tying compensation to financial measures such as reported earnings per share, revenues, stock prices, and return on equity.

Prior to the 1970s, only 16 percent of the chief executive officers in S&P 500 companies had performance-based compensation, but this proportion grew to 26 percent in the 1980s and 47 percent in the 1990s (Bank, Cheffins, and Wells 2017). The vast majority of large corporations today use earnings per share in incentive plans, and most use stock prices and shareholder returns in their compensation plans (Reda, Schmidt, and Glass 2016). Compensation for managers (as well for as boards) typically includes restricted stocks and options. In this way, corporations are effectively "managed" by markets and by accounting-based metrics (Davis 2011).[1]

[1] An alternative approach to motivating managers to focus on shareholder value relies on the market for corporate control (Manne 1965). The idea is that firms whose managers do not maximize shareholder value as measured by the stock price will be targets of hostile takeovers and the underperforming managers will be replaced. However, boards and managers can find ways to raise the costs of hostile takeovers such as poison pill provisions, and governments may block takeover transactions because of political pressures. Most corporate mergers today are "friendly."

The prevalence of stock-based compensation affects the efficacy of corporate governance arrangements, but understanding the issues around corporate governance more fully requires a broader context. First, the shareholders of most public corporations today are not individuals, but rather institutional investors such as mutual funds, pension funds, hedge funds, or endowments, which are usually corporations themselves with their own governance challenges. Second, corporations may set up and invest in corporate subsidiaries, creating complex corporate structures. In this environment, stock prices do not measure properly whether managers actually benefit the majority of their ultimate shareholders. Third, some of the tradeoffs associated with financialized corporate governance are relevant even in the absence of shareholder–manager conflict and arise in the context of private corporations as well.

Consider the layered ownership structure of public corporations. Institutional investors accounted for only 6.1 percent of corporate ownership in the 1950s, and, by 2009, this fraction grew to 73 percent for the top 1,000 largest US corporations (Gilson and Gordon 2013). Mutual funds are usually subsidiaries of "management companies," which are separate corporations with their own objectives (Bogle 2005). This ownership system creates new agency problems between corporate managers in the firms along the ownership chains and the investors at the ends of the chains. Moreover, those who control institutional investors have their own objectives that may conflict with their clients. The managers of institutional investors often have little incentive to engage in the governance of portfolio firms even if it would benefit ultimate investors (Taub 2009). Gilson and Gordon (2013) refer to the conflicts between the interests of funds' managers and investors as the "agency costs of agency capitalism."

Even if individuals held corporate shares directly, it is unclear that maximizing "shareholder value" as currently practiced captures the preferences of most or all shareholders. First, high-powered financialized incentives may be counterproductive when managers have multiple tasks (Holmstrom and Milgrom 1991). Second, modern portfolio theory suggests that investors should diversify their holdings, which means that shareholders often own shares in multiple firms in the same industry. As shareholders, they may benefit if firms collude, but lack of competition harms them as customers or employees and distorts the economy. Indeed, shareholder unanimity is not assured except under unrealistic assumptions such as complete markets and perfect competition. Third, the ability to engage in short selling and trade derivatives can decouple the economic interests of some shareholders from their voting rights (Barry, Hatfield, and Kominers 2013).

An interesting phenomenon in a broader corporate governance context is the proliferation of opaque shell corporations with no employees or publicly traded shares (Story and Soul 2015). Individuals and corporations often create them to limit liability, hide activities, or avoid taxes or other laws. Many jurisdictions, including Delaware (the most popular US state for incorporation) do not require any information about the shareholders—the so-called "beneficial owners"—of corporations they register. One office building in Delaware is the legal address of 285,000 separate businesses; Delaware uses revenues from taxes and fees by absentee corporations to

fund a significant part of its budget, and it has fought against federal legislation that would increase the transparency of corporate ownership (Wayne 2012).

Tradeoffs from Financialized Corporate Governance

Financialized, shareholder-focused governance is appealing in its logic. However, in addition to the issues already raised above, it introduces tradeoffs and potential distortions that can have significant effects on the economy. Corporations interact with most of their stakeholders, other than shareholders, through contracts and markets. Counterparties will be more willing to engage with corporations, make investments, and produce economic efficiencies if they trust that corporations would not harm them subsequent to their investments (Mayer 2013). For example, if lenders cannot trust the legal system to collect loans in a timely manner or prevent borrowers from exposing them to additional risk once the loan is made, they will refuse to make loans or charge a high rate of interest. Creating trust requires being able to make credible commitments, but making commitments may be impossible, difficult, or costly. Dealing with externalities may require government action.

The cost of making and enforcing commitments is ultimately borne by the corporations' residual claimants and by society as a whole through the government that creates and enforces the rules. For corporations and their governance to support the economy best, it is important that contract enforcement be efficient, markets be competitive, and appropriate rules correct market failures and externalities. Financialized governance aims to focus corporate managers on benefitting shareholders, but it can result in gaps between what is good for executives, directors, and some shareholders and what is good for society as a whole.

I will focus on two types of tradeoffs that derive from frictions such as asymmetric information and the difficulty and cost of effective commitments. First, financialized governance may lead managers to manipulate disclosures and engage in deception, fraud, or other misconduct. Second, financialized governance may cause inefficiencies through misallocation of resources and risk. The culprit in many of the examples appears to be a focus on financial metrics. The inefficiencies ultimately link to the weak or lacking incentives of those who are in a position to put in place mechanisms to prevent harmful conduct.

Corporate Opacity, Fraud, and Deception

Enforceable contracts and effective governance require reliable and verifiable information. Extreme information asymmetries can cause markets, contracts, laws, and the potential discipline of reputation concerns to break down. Thus, providing information that enables markets and contracts to function well, and which allows effective control and accountability, is a key governance issue.

Managers whose compensation depends on financial targets have incentives to divert time and energy to actions that improve the appearance of meeting or exceeding short-term financial targets. For example, managers may engage in

"managing" earnings within allowable accounting standards (Teoh, Welch, and Wong 1998; Graham, Harvey, and Rajgopal 2005). These activities may become deceptive or fraudulent, as happened at Tyco, Enron, WorldCom, and numerous other institutions. Complex transactions in opaque derivatives markets and the creation of off-balance-sheet subsidiaries make it difficult to detect or distinguish financial fraud from other misleading disclosures, as illustrated by Lehman Brothers' use of "repo 105" transactions (Eisinger 2017). The complexities of securitization and derivatives allow banks to manipulate valuations and hide losses (Piskorski, Seru, and Witkin 2015). Opaque off-balance-sheet subsidiaries can make large banking institutions appear as "black boxes" to investors (Partnoy and Eisinger 2013).

Corporate fraud or misrepresentation can remain hidden for extended periods or even indefinitely (Zingales 2015), which prevents effective accountability. It is often hard to pin the responsibility and intent to specific and appropriate individuals. There are also insufficient incentives or willingness within corporations to uncover fraud or deception, particularly if executives are able to benefit from such practices. Whistleblowers face hardships, lose jobs and opportunities, and may be unable to prevail if authorities are not inclined to pursue their claims (Sawyer, Johnson, and Holub 2010; Ben-Artzi 2016). Even if it is possible to trace misconduct to specific individuals, markets may do little to correct the problem. Financial advisors with records of misconduct continue to find employment (Egan, Matvos, and Seru 2017).

The problem extends to auditors, which are supposed to be independent watchdogs, but in fact have weak incentives to uncover fraud and do not opine on the absence of fraud. Despite accounting scandals in the early 2000s that led to attempts to improve the quality of audits in the United States, Ronen (2010, in this journal) describes auditors as "lapdogs" and the *Economist* (2014b) calls them "dozy watchdogs." Four large, for-profit corporations with little accountability to the public dominate the auditing industry. These companies are opaque themselves, and some, such as KPMG, have been accused of fraud and obstruction of justice repeatedly in recent years (Eisinger 2017).

Consumer fraud or deception, and other law evasion or misconduct, may actually benefit shareholders, particularly if the misconduct remains hidden. Of course, if and when problems come to light, the legal costs, fines, and loss of reputation affect the corporation's success and are borne by shareholders, employees, and possibly others. Recent examples include Volkswagen's evasion of environmental regulations and the case of Wells Fargo Bank "cross selling" and improperly opening accounts. New information on corporate prosecutions and misconduct keeps coming to the surface.[2]

[2]A new Corporate Prosecution Registry (Garrett and Ashley 2017) at the University of Virginia Law School collects data on corporate prosecutions (at http://lib.law.virginia.edu/Garrett/corporate-prosecution-registry/index.html). The nonprofit Corporate Research Project collects information with the purpose of increasing corporate accountability, including "corporate rap sheets" (at http://www.corp-research.org/).

The costs to society of corporate opacity, fraud, and deception are high. Lack of trust by shareholders and other investors can increase the funding costs of corporations. Lenders who fail to recognize loan losses may avoid restructuring loans and continue to lend to insolvent borrowers rather than making new loans. Lingering debt overhang for households and lenders can contribute to long-term recessions that harm entire economies, as happened in Japan in the 1980s, in the United States during the housing crisis, and European nations today (Admati and Hellwig 2013; Mian and Sufi 2015). Ownership chains involving shell corporations can also enable fraud and make contract enforcement and beneficial renegotiation more difficult, all of which were evident in the recent mortgage crisis (Dayen 2016).

More subtle and harder to address are corporate strategies involving systematic and harmful deception that may cause significant social harm to shareholders and consumers. Consider, for example, tobacco companies that denied the addictiveness and harm from cigarettes for decades even as they had information inconsistent with the claims they made, or the campaign by the sugar industry to distort nutrition research and dietary guidelines by diverting attention away from the harm of sugar consumption. Akerlof and Shiller (2016) discuss these and other cases where manipulation and deception by profit-maximizing corporations have caused distortions and harm. The main weapon against such strategies is public education and awareness of how conflicts of interests can corrupt information sources, including even supposedly neutral academic research.

Misallocation of Resources and Risk

A related but somewhat different set of tradeoffs from financialized corporate governance involve inefficiencies from misallocation of resources and risk. First, managers may display "short-termism" in response to short-term accounting metrics and pass up worthy long-term investments (Graham, Harvey, and Rajgopal 2005). Second, financialized governance can encourage managers to endanger stakeholders—for example, by compromising product quality, the health and safety of customers or employees, or even the solvency of the corporation—particularly if such actions remain hidden and still allow the manager to be rewarded upfront, before risks materialize. Shareholders may be harmed by being exposed to excessive risks without compensation or even knowledge of the risk, but sometimes they benefit from endangering or harming other stakeholders.

Because stock prices reflect assessments of future cash flows, stock-based compensation is less prone to causing distortions than compensation based on short-term accounting measures. In theory, if all investors have the same information as managers, their holding periods or investment horizons do not matter, and neither does the timing of dividends. In that special case, the stock price reflects the consequences of all corporate action for shareholders. If managers of public corporations reinvest profits in worthy projects, shareholders who need immediate cash can sell shares at prices that reflect the investments.

Accordingly, in the standard teaching of basic finance, shareholders agree that managers should invest in projects that create value for the corporation, and

increases in firm value raise the share price. The conclusions change if managers have different information than investors. In such cases, managers may make inefficient decisions that harm shareholders (and possibly others) while inflating stock price even in the absence of an underlying managers–shareholders conflict (Narayanan 1985; Stein 1989).

Compensation based on earnings or return on equity targets without accounting for risk creates significant distortions that can harm shareholders (Admati and Hellwig 2013). For example, it encourages managers to magnify risk by using debt even if doing so harms shareholders and others. The incentives are particularly strong if managers can reduce taxes for the corporation or take on risk in ways that magnify the upside for shareholders while sharing downside risk with others.

Managers can also "front load" the upside and reap large bonuses, because return measures are high at first while potential losses, realized later, fall mainly on shareholders and others (Bhagat 2017). Those who manage institutional investors such as asset management companies, pension funds, mutual funds, and endowments may also be judged by short-term return measures and expose the ultimate investors to excessive risk (Bogle 2005; Partnoy 2009). In some cases like public pension funds, banks with insured deposits, or institutions whose creditors are likely to receive support from governments or central banks, a share of the downside risk ultimately falls on taxpayers.

Risk taking in innovation, where those who take the risk bear the downside, is useful and beneficial if taken properly and responsibly. Indeed, managers, fearing for their jobs, may be excessively risk averse and take too little such risk. The problem of excessive risk taking arises when executives can shift downside risk and endanger others inefficiently. Cases such as Volkswagen, British Petroleum, or the nuclear industry in Japan illustrate the problem and the potential harm that can result. Dispersed consumers or the public do not have sufficient information or ability to bring about safer practices or to prompt action to eliminate products that turn out to be unsafe (Fletcher 2001).

Another example of the harmful consequences of financialized corporate governance that may lead to lower firm value and collateral harm is excessive use of debt funding by corporations. Managers acting on behalf of shareholders of indebted corporations make investment and funding decisions that may not maximize the total value of the corporation. In particular, they may make excessively risky investments and increase indebtedness inefficiently because shareholders benefit fully from the upside of risk while sharing an increased downside risk with creditors (or others). At the same time, indebted corporations avoid taking actions that benefit creditors and the corporation as a whole at shareholders' expense, such as beneficial reductions of indebtedness and some worthy investments that do not have sufficient "upside" potential for shareholders (Admati, DeMarzo, Hellwig, and Pfleiderer forthcoming).

Heavy borrowing thus leads to distorted investments and to an increased risk of defaults and bankruptcies that entail deadweight cost and, for large corporations, can cause collateral harm to employees, customers, and the community. The

problem of excessive and reckless use of debt is particularly harmful in banking, where passive depositors and short-term creditors do not provide market discipline, and explicit and implicit guarantees exacerbate the distortions, essentially feeding a "debt addiction" that characterizes heavy borrowing. Unless regulations counter the harmful incentives, the result is distorted credit markets; financial instability, including periodic financial crises; and further governance problems, recklessness, and distorted competition when institutions are considered "too big to fail."

Some Policy Proposals

The key to improving corporate governance is to increase transparency, create better internal and external control and accountability, and address distortions and inefficiencies through effective laws and regulations. With financialized governance, executives will obviously seek to maintain market power and prevent entry, and antitrust laws should attempt to promote competition and entry. I will focus on addressing the potential inefficiencies from opacity, fraud, and excessive endangerment discussed above.

One place to start reducing corporate opacity would be to require shell corporations to reveal the identity of their beneficial owners, and any limits to their liability, so that authorities and the public can better track chains of ownership. Such laws exist in many jurisdictions but, surprisingly, not in the United States (Caldwell 2016). It also makes sense to consider whether the privilege of incorporation should be available as easily as it is now. One idea is that incorporations would require a disclosure of purpose, at least in general terms, which would be revised and examined periodically with possible termination if the corporation is primarily set for the purpose of increasing opacity and evading laws. Such examinations could also lead to charges of tax evasion or fraud.

For large corporations, it may be useful to find more unconflicted sources of information outside the corporations by providing incentives to independent analysts to expose misconduct, given the difficulty of relying on whistleblowers and the conflicts of interest of auditors and rating agencies paid by the corporations. Since producing reliable information is so critical for effective governance, it may be desirable to delegate some of these functions to government agencies or to not-for-profit organizations with committed and unconflicted experts.[3] Unless rating agencies are more accountable to the public, regulations and institutional investors should avoid relying on their scores (Partnoy 2016).

As abstract entities, corporations cannot go to jail. Extracting fines from corporations does not prevent corporate fraud and misconduct if shareholder governance is weak. The individuals who are involved in, encourage, or tolerate

[3] Shifting the responsibility for choosing auditors to private insurance companies (Ronen 2010) may be helpful, but it does not address the distorted incentives of individuals in response to their own compensation and the lack of personal accountability when responsibility is diffused.

corporate misconduct or law evasion often benefit from effective personal impunity because their personal culpability or intent cannot be established with sufficient confidence to meet a legal standard. Unless shareholder governance is effective, corporate misconduct rarely leads to significant negative personal consequences for executives and board members.

The ability to deter large corporations from bad behavior is limited by the fact that imposing the most severe punishments—huge fines, or worse, the revocation of license to conduct business—would cause significant collateral harm to innocent employees and others (Garrett 2016). Such issues do not arise if we increase accountability for individual executives and board members. Doing so may require re-examination of the laws and rules defining liability that would give authorities sufficient tools to pursue individuals in civil and criminal courts, and to claw-back pay. Devoting sufficient resources to investigations of individuals, which tend to be complex and risky, may also be necessary (Eisinger 2017).

There have been attempts to improve corporate governance and prevent accounting fraud through laws. However, the Sarbanes–Oxley Act of 2002 that came as a response to the Enron bankruptcy and the numerous accounting scandals around that time did not prevent the massive fraud and deception by many financial firms that contributed to the housing crisis and to the near implosion of the financial system in 2008 (Coates and Srinivasan 2014). There is also no evidence that independent directors have prevented fraud (Avci, Schipani, and Seyhun 2017). The 2010 Dodd–Frank Act has done little to address corporate fraud except for attempting to encourage whistleblowers.

Many deceptive practices fall in a gray area where it is difficult to identify or establish that they are fraudulent with intent to deceive as defined under law. To prevent corporations from hiding safety problems of which they are aware, laws are needed to force corporations to take strong action to inform consumers about safety issues and to prohibit settlements that specifically obscure safety violations. Consumer protection laws are useful when it is difficult for consumers to evaluate products—for example, in the context of financial services (Campbell 2016). Educating the public to be more aware of potential conflicts of interest, thus creating savvier consumers of products and information, including from experts and media, would also help.

To address the problem of corporations transferring risk inefficiently to others and misallocating resources, it is important that incentives offered to managers create a long-term focus. Corporations should also have processes to ensure that relevant information about safety issues is not diffused or lost and reaches executives in positions of control. Measures that prevent or reduce harm are obviously better for all, including shareholders who would otherwise deal with fines and the company's loss of reputation.

Effective laws and regulations are essential when competitive markets and contracts do not work to create effective commitments or there are externalities. In creating laws and regulations, the key should be first on prevention of harm if it can be achieved at a reasonable cost, rather than focusing on how to deal

with the conduct after the fact. For example, preventing traffic accidents through appropriate traffic laws such as speed limits and proper infrastructure is better than relying solely on insurance, fines, prisons, civil litigations, and ambulances. Similarly, it may be significantly more cost efficient and prevent collateral harm to try to detect and address misconduct, fraud, and endangerment early than to deal with consequences such as nuclear disasters, oil spills, car explosions, or financial crises once they happen. In the case of children's products in the United States, for example, safety standards are lax and corporations often obscure information about unsafe products (Felcher 2001).

Of course, it is important that policymakers choose the least costly ways to achieve prudent conduct. Yet, some laws are counterproductive and interfere with efficient corporate governance. For example, tax laws in many jurisdictions favor debt over equity funding. Such laws are distortive by creating incentives for inefficient indebtedness (Hirshleifer and Teoh 2009; Admati et al. forthcoming). This feature of tax codes is particularly perverse for banks, which already have incentives to choose dangerous debt levels. The *Economist* (2015) called tax-free debt "a vast distortion in the world economy [that] is wholly man-made." Bankruptcy codes that favor commitments in derivatives and short-term debt (so-called repos) over other corporate liabilities, and which also exacerbate the conflict of interest between managers with financialized compensation and society, should be changed (Skeel and Jackson 2012).

Political Economy and Corporate Governance

By putting in place laws and regulations and by enforcing contracts and rules, governments play a critical role in affecting corporate governance practices and determining how well corporations serve society. The determination of the rules, and how they affect different stakeholders, in turn depends on policymakers' incentives and on the political process (Pagano and Volpin 2005). Policymakers may help corporations create useful commitments and thus become more efficient, or instead impose excessive and costly rules on some corporations while tolerating or even perversely encouraging reckless conduct in other contexts.

To see some of the issues, it is instructive to compare corporate governance and aviation safety. A key reason for the safety of airplane travel is that lapses in safety are extremely salient to the public. Authorities design rules that anticipate and reduce potential problems, and they investigate problems promptly. In addition, the incentives of those in the private aviation sector, from the airplane manufacturers to the airlines employees to those working in airports to monitor air traffic, do not conflict with the public's interest in safety. Finally, a key underlying reason for aviation safety has to do with accountability. In virtually all plane crashes, it is possible to point to the cause of the crash. Individuals found responsible or negligent stand to lose jobs or reputation from plane crashes, and they might even get into legal trouble. Although it takes much technology and collaboration across jurisdictions, safety prevails in aviation and mistakes rarely recur.

Corporate governance issues are in some cases starkly different. When those in control of corporations can harm others in abstract or invisible ways, through excessive financial risk or other subtle endangerment, governments may lack the *political will* to consider the issues, do a thorough autopsy when problems arise, or invest properly in putting in place effective rules to prevent the problems from repeating. Instead, governments may enact inefficient, excessive, or wasteful rules that create or exacerbate distortions in order to serve other political objectives.

Even when corporate governance failures become clear, for example in scandals or crises, it is often hard to trace the harm to specific individuals or policies. The governance and accountability of government institutions can become a challenge for society. In this section, I discuss several issues that arise at the intersection of political economy and corporate governance: capture, law enforcement, and companies operating across legal jurisdictions. In the next section, I offer the financial industry as an example in which these issues are particularly stark.

Capture

Laws and regulations will not work well when those charged with setting and implementing them collaborate with those in the industry even if these collaborations harm the public (Stigler 1971; Acemoglu 2003). The dynamics of capture are often subtle. Corporations employ lobbyists, consultants, lawyers, public relations firms, and influential, connected individuals to shape rules and their implementation. Such activities have expanded greatly in recent years (Drutman 2015). The realities of revolving doors and campaign finance in the United States have increased the impact of those who can fund politicians (Lessig 2012).

When the issues are complex and government resources are limited, staffers and policymakers sometimes rely on corporations and their lobbyists to draft rules (Lipton and Protess 2013). Complex laws and regulations create a bloated ecosystem of experts who find revolving opportunities in the private and government sector based on knowing the relevant details (McCarty 2013; Lucca, Seru, and Trebbi 2014).

The actual workings of capture and the corrosive impact it can have on the effectiveness of governments are often invisible. If budgets are tight and expertise lies mostly with conflicted individuals, rules are more likely to become distorted and fail to serve the public interest. The "thin political markets" that produce the rules do not balance the interests of different constituents, affecting even basic accounting rules, which are the fundamental building blocks of effective governance (Ramanna 2015). The mix of genuine confusion and distorted incentives compounds the problems and leads to "intellectual capture" (Johnson and Kwak 2010).

Given the critical importance of appropriate and well-crafted rules, reducing the wage disparity between policymakers and the private sector would be desirable. Low salaries encourage the government-to-lobbyist revolving door and may deprive the government of experts who are more likely to stand up to pressure from the industry and protect the public interest through effective rules (Drutman 2015).

Corporations fight against rules and their implementation in courts, where outcomes often depend on the biases and ability of specific judges to understand the complex issues and on the quality of the lawyers making the arguments. The resources of corporations often overwhelm those that governments are able or willing to devote to the issues.

It does not follow from this discussion of capture that governments should impose no rules on corporations or, alternatively, that all regulations are useful. Rather, my point is that the incentives of those who work in government matter and that it is important that they use their power properly and be accountable to the public. Governments can fail by intervening too much or too little, by creating inefficient and excessively complex rules, or by not devoting enough resources to writing and enforcing rules. Rules should be as cost-beneficial as possible to address market failures while avoiding waste of taxpayer or corporate resources. Preventing capture and providing proper incentives for regulators and others involved in policy is itself an important objective (Carpenter and Moss 2013).

Effectiveness of Enforcement

A related issue is that laws and regulations may fail to achieve their goals if governments do not enforce them consistently and effectively. As a representative example, consider the Deutsche Bank whistleblower who contacted the Securities and Exchange Commission (SEC) to report a significant mismarking of derivatives positions; this case only received attention after the media investigated and reported the allegations (Ben-Artzi 2016). The result was a fine of $55 million, effectively paid by current shareholders, with little if any direct consequences for those responsible for the fraud. Revolving doors between Deutsche Bank compliance and SEC enforcement may have played a role in this case.

The US Department of Justice and other regulatory agencies have changed how they handle corporate crime, particularly fraud, since the late 1990s. The main tool has become settlements with deferred prosecutions and fines, while indictments of individuals, particularly executives, have become extremely rare since the cases of Enron and others in the early 2000s. Among the reasons for the shift is the length and complexity of investigations and trials of individuals, lack of investigative resources, and the loss of some legal tools to pursue individuals (Eisinger 2017). However, large fines do not appear to change corporate culture or act as a deterrent (Garrett 2016).

If lack of resources undermines enforcement, misconduct is even less likely to surface, and thus it can become more prevalent. For example, the 2010 Dodd–Frank Act expanded the scope of the Commodities and Futures Trading Commission (CFTC)'s jurisdiction dramatically, beyond the $34 trillion US futures market to the much larger market in derivatives traded outside established exchanges estimated to be as large as $400 trillion in so-called "notional value." Yet the agency is severely underfunded relative to other agencies and given the enormous size of the markets it oversees. One person at the CFTC oversees the $117 billion US market where wholesale prices for gasoline and heating oil are set (Leising 2017). The departing head of compliance of CFTC said in March 2017 that the agency is unable

to investigate the "massive amount of misconduct" in derivatives markets (Freifeld 2017). The effectiveness of banking regulations also depends significantly on the resources and incentives of regulators (Agrawal, Lucca, Seru, and Trebbi 2014).

Regulation across Jurisdictions

The political economy of corporations involves competition among jurisdictions. This competition can happen within countries: as noted, state-level corporate havens such as Delaware may benefit while harming taxpayers and citizens in other jurisdictions such as the US federal government. Holding corporations responsible can be even harder in the context of a global economy. At the international level, Panama, Liberia, and Bermuda are popular havens for many corporations and wealthy individuals (Davis 2011), but the United States and some other developed nations are among the easiest places to hide wealth (*Economist* 2016).

Corporations can "shop jurisdictions" and set up opaque corporations or subsidiaries that allow them to avoid taxes or other laws (OECD 2015). The process of negotiating and coordinating international regulation often results in a race to the bottom that lessens the effectiveness of the regulations that would have otherwise been adopted in at least some countries. Politicians tend to side with "their" corporations, because corporate voice is more salient to them than the broader and more passive public whose voice might be missing (Admati and Hellwig 2013, chap. 12).

Corporations have also used international trade agreements to challenge actions of governments. Opaque tribunals of private lawyers, where corporations can sue but governments cannot sue or appeal on behalf of their citizens, adjudicate disputes between corporations and national governments (*Economist* 2014a).

Corporate Governance in the Financial Sector

Banks and the financial industry provide an extreme illustration of the distortions created by financialized corporate governance and the shortcomings of laws and regulations. History shows that in the context of banking, governments often lack the political will needed to address market failures, and the difficulty of commitments, by means of effective rules. Sovereign default and other government actions have often caused banking crises (Reinhart and Rogoff 2009).

Today, and even after the crisis of 2007–2009, the result of the combined failure of corporate governance and policy is a set of overly fragile financial institutions and a highly interconnected and fragile system that endangers the economy unnecessarily. In extreme contrast with aviation, where many individuals and institutions collaborate to maintain safety, most of those within the private and public institutions involved in the financial sector benefit personally from practices that create excessive endangerment and that conceal this reality from the public (Admati and Hellwig 2013; Admati 2017).

Economists treat banks as special because of their role in the payment system and their intermediation function, although loans can be—and are—made by other types of institutions. Because banking has always been fragile and has repeatedly produced cycles of booms, busts, and crises, a common view is that fragility is inherent to banking and fundamentally unavoidable.

It is true that banks are prone to liquidity problems: that is, circumstances can arise in which they have trouble converting illiquid assets to cash quickly at a reasonable price to satisfy creditors' demands. These problems can result in panics and runs if depositors and short-term creditors withdraw their funding. Banks can reduce the likelihood of such problems by reducing their opacity and indebtedness (for example by using their profits as a source of funding or issuing more shares and having better disclosures). However, banks have been able to remain dangerously and inefficiently indebted and to obscure the true exposure to risk of their shareholders, creditors, and taxpayers through opaque disclosures.

When banks were run as partnerships in 19th century England, they commonly funded half of their loans with equity, and their owners or shareholders had unlimited liability, exposing their personal wealth to the risk that their bank's assets would not be sufficient to pay deposits. A century ago in the United States, bank equity levels were around 20 percent or more and shareholders often had increased liability. Over the years, banks became limited liability corporations, and some operate within large holding companies engaging in extensive trading and other activities beyond making loans to individuals and businesses. To prevent disruptions from liquidity problems and runs, governments have created safety nets such as deposit insurance and central bank lending. These safety nets weaken and can even lead to the breakdown of corporate governance.

What actually makes banks and other financial institutions "special" is their unusual ability to shift downside risk and costs to others and the fact that normal market forces do not work to counter the distorted incentives of those who control them. For example, outside banking, bankruptcy courts prevent shareholders of insolvent corporations from benefiting at the expense of creditors, for example, by "looting" the corporation or gambling for resurrection inefficiently. By contrast, hidden insolvencies can persist in so-called "zombie banks" if authorities do not intervene, because depositors and short-term creditors use their ability to withdraw funding, close out their positions, or count on explicit or implict guarantees to protect themselves (Akerlof and Romer 1993; Skeel and Jackson 2012).

Financial innovations such as securitization and derivatives, and the creation of complex structures around the globe, have also allowed financial institutions to take risks and increase their indebtedness while hiding their true financial health from investors and regulators (Partnoy and Eisinger 2013). Corporate structures are particularly complex and opaque in large banking institutions (Carmassi and Herring 2014).

Poor risk governance and the distorted incentives of traders, described in many books about the culture of banking since the 1980s (for example, Partnoy 2009; Das 2010), appear to persist. The US Senate investigation of the JPMorgan

Chase "whale trades" in 2013, which involved taking huge positions in thinly traded markets in London, leading to losses of over $6 billion, showed that risk controls in at least some of the largest institutions remain highly problematic (Norris 2013). But except in such extreme cases, or after bankruptcies or crises, poor risk governance in banking is invisible.

Governments can counter the incentives for endangerment in banking, for example, by insisting that shareholders bear more of the risks they take and by reducing the opacity of the system through better disclosures and tracking of risk. Bank lobbyists often threaten that such steps would "harm credit and growth." In fact, the most costly and harmful outcomes arise from a combination of too much credit in boom times, overly complex and ineffective regulations that exacerbate governance and other distortions, and "extend and pretend" policies that tolerate and support insolvent and dysfunctional banks and other borrowers for too long.

The dynamics of regulatory capture are particularly strong in the financial sector (Connaughton 2012). US Senator Richard Durbin admitted in a 2009 interview that "banks are still the most powerful lobby in Capitol Hill and they frankly own the place."

The regulatory capture problem arises because politicians often view banks and financial firms as a source of funding for favored projects rather than as a source of risk for the public, and thus choose to cut deals that compromise efficiency and stability. Even after the devastating financial crisis of 2007–2009 and the recession that followed, policymakers failed to learn key lessons. Implicit guarantees, which perversely encourage and reward recklessness and are ultimately costly to the public, appear free to politicians. The jargon and technical issues and the abstract nature of the risk muddle the policy debate and create public confusion about the issues and the relevant tradeoffs (Admati and Hellwig 2013; Admati 2016, 2017).

Other misconduct such as fraud and deception plague the financial sector, leading to invisible harm to many and to hundreds of billions in fines in recent years (Zingales 2015). The largest financial institutions, considered "too big to fail," have outsized power that distorts competition and the economy, and they are especially inefficient and dangerous being effectively above the rules. Fragmented regulatory structures, such as in the United States, and the ability to play off governments and regulatory agencies have made financial regulation particularly challenging to design and enforce. The main problem remains the lack of collective *political will* to create a safe and efficient global financial system.

Conclusion

Milton Friedman (1970) famously argued that the social responsibility of corporate managers is to "make as much money as possible while conforming to the basic rules of society, both those embodied in law and those embodied in ethical custom." Friedman presumes that the firms operate in an environment of "open and free competition without deception and fraud," and he warns that chief executive

officers who "pontificate" about corporate social responsibility will bring back "the iron fist of government bureaucrats."

However, "free and open" markets will not necessarily become competitive and devoid of deception and fraud on their own. Rules matter. The limited liability and separate legal status of corporations have benefits but also create problems of misaligned incentives, and lack of individual accountability exacerbates these problems. Those who manage firms will respond in predictable ways to financialized incentives. Private sector mechanisms such as auditors or rating agencies are unlikely to uncover fraud, or provide reliable information, without law enforcement and proper regulations and oversight.

The interactions between governments and corporations can promote efficiency, but even in well-functioning democracies, they can also be wasteful and exacerbate distortions that benefit only a few. The issue is not the *size* of governments, but rather conflicts of interests affecting people in all institutions, and particularly the quality, integrity, and effectiveness of the institutions that design, implement, and enforce the rules.

Distortions from inefficient corporate governance are important determinants of economic outcomes. To ensure competition and create accountability, brave and well-informed policymakers—including brave bureaucrats—must erect and implement effective systems that can counter the incentives of corporate managers to extract rents, deceive, and mismanage risk. In a democracy, individuals in government must also be accountable if they fail to act in the public interest. In reality, inefficient governance may persist.

The status quo, in which governments too often tolerate or exacerbate corporate governance distortions rather than alleviate them, is dangerous and harmful. Positive change requires better understanding of the underlying causes. Economists can play an important role by studying these important issues, clarifying the tradeoffs associated with governance mechanisms, identifying instances where markets and institutions cause harm, and suggesting approaches to reduce the scope for abuses of power by individuals in all institutions. Increasing transparency, holding those in control more accountable, and creating and enforcing better laws and regulations to address corporate fraud and endangerment would be highly beneficial.

■ *I am grateful to Daron Acemoglu, Jon Bendor, Anne Beyer, Steve Callander, Peter Conti-Brown, Lee Drutman, Gordon Hansen, Martin Hellwig, David Hirshleifer, Peter Koudijs, David Kreps, Signe Krogstrup, Enrico Moretti, Kjell Nyborg, Saule Omarova, Frank Partnoy, Jeff Pfeffer, Paul Pfleiderer, Elizabeth Pollman, Heiner Schulz, Amit Seru, Eytan Sheshinski, Sarah Soule, Jennifer Taub, Timothy Taylor, and Jeff Zwiebel for very helpful discussions and comments and to Nathan Atkinson, Andrew Baker, Zhao Li, and Sara Malik for excellent research assistantship.*

References

Acemoglu, Daron. 2003. "Why Not a Political Coase Theorem? Social Conflict, Commitment, and Politics." *Journal of Comparative Economics* 31(4): 620–52.

Admati, Anat R. 2016. "The Missed Opportunity and Challenge of Capital Regulation." *National Institute Economic Review* 235(1): R4–14. (Related materials linked at https://www.gsb.stanford.edu/faculty-research/excessive-leverage.)

Admati, Anat R. 2017. "It Takes a Village to Maintain a Dangerous Financial System." Chapter 13 in *Just Financial Markets? Finance in a Just Society*, edited by Lisa Herzor. Oxford University Press.

Admati, Anat R., Peter M. DeMarzo, Martin F. Hellwig, and Paul Pfleiderer. Forthcoming. "The Leverage Ratchet Effect." *Journal of Finance.*

Admati, Anat R., and Martin F. Hellwig. 2013. *The Bankers' New Clothes: What's Wrong with Banking and What to Do about It.* Princeton University Press. (Excerpts and more materials linked at http://bankersnewclothes.com/.)

Agrawal, Sumit, David Lucca, Amit Seru, and Francesco Trebbi. 2014. "Inconsistent Regulators: Evidence from Banking." *Quarterly Journal of Economics* 129(2): 889–938.

Akerlof, George A., and Paul M. Romer. 1993. "Looting: The Economic Underworld of Bankruptcy for Profit." *Brookings Papers on Economic Activity* 2: 1–73.

Akerlof, George A., and Robert J. Shiller. 2016. *Phishing for Phools: The Economics of Manipulation and Deception.* Princeton University Press.

Avci, Burcu S., Cindy A. Schipani, and H. Nejat Seyhun. 2017. "Do Independent Directors Curb Financial Fraud? The Evidence and Proposals for Further Reform." Ross School of Business Working Paper 1352.

Bank, Stephen A., Bran R. Cheffins, and Harwell Wells. 2017. "Executive Pay: What Worked?" UCLA School of Law Law-Econ Research Paper 16–11.

Barry, Jordan M., John William Hatfield, and Scott Duke Kominers. 2013. "On Derivatives Markets and Social Welfare: A Theory of Empty Voting and Hidden Ownership." *Virginia Law Review* 99(6): 1103–67.

Bebchuk, Lucian A., and Michael S. Weisbach. 2010. "The State of Corporate Governance Research." *Review of Financial Studies* 23(3): 939–61.

Ben-Artzi, Eric. 2016. "We Must Protect Shareholders from Executive Wrongdoing." *Financial Times*, August 18. https://www.ft.com/content/b43d2d96-652a-11e6-8310-ecf0bddad227.

Bhagat, Sanjai. 2017. *Financial Crisis, Corporate Governance, and Bank Capital.* Cambridge University Press.

Bogle, John C. 2005. *The Battle for the Soul of Capitalism.* New Haven, CT: Yale University Press.

Caldwell, Leslie. 2016. "End the Corporate Shell Games." *Bloomberg View*, November 30. https://www.bloomberg.com/view/articles/2016-11-30/end-the-corporate-shell-games.

Campbell, John Y. 2016. "Restoring Rational Choice: The Challenge of Consumer Financial Regulation." *American Economic Review* 106(5): 1–30.

Carmassi, Jacopo, and Richard J. Herring. 2014. "Corporate Structures, Transparency and Resolvability of Global Systemically Important Banks." https://fic.wharton.upenn.edu/wp-content/uploads/2016/11/15-10.pdf.

Carpenter, Daniel, and David Moss. 2013. *Preventing Regulatory Capture: Special Interest Influence and How to Limit It.* Cambridge University.

Cecchetti, Stephen G., and Enisse Kharroubi. 2015. "Why Does Financial Sector Growth Crowd Out Real Economic Growth?" BIS Working Paper 490.

Coates, John C., and Suraj Srinivasan. 2014. "SOX after Ten Years: A Multidisciplinary Review." *Accounting Horizons* 28(3): 627–71.

Connaughton, Jeff. 2012. *Payoff: Why Wall Street Often Wins.* Westport, CT: Prospecta Press.

Cournéde, Boris, and Oliver Denk. 2015. *Finance and Economic Growth in OECD and G20 Countries.* OECD Economics Department Working Paper 1223.

Das, Satyajit. 2010. *Traders, Guns and Money: Knowns and Unknowns in the Dazzling World of Derivatives.* 2nd edition. London: Financial Times (Prentice Hall).

Davis, Gerald F. 2011. *Managed by the Markets: How Finance Re-shaped America.* Oxford University Press.

Dayen, David. 2016. *Chain of Title: How Three Ordinary Americans Uncovered Wall Street's Great Foreclosure Fraud.* New Press.

de Haan, Jacob, and Jan-Egbert Strum. 2016. "Finance and Income Inequality: A Review and New Evidence." De Nederlandsche Bank Working Paper 530.

Drutman, Lee. 2015. *The Business of America is Lobbying: How Corporations Became Politicized and Politics Became More Corporate.* Oxford University Press.

Durbin, Dick. 2009. Interview by Bill Moyers for *Bill Moyers Journal*, May 8. http://www.pbs.org/moyers/journal/05082009/transcript1.html.

Economist, The. 2014a. "The Arbitration Game," October 11. http://www.economist.com/news/finance-and-economics/21623756-governments-are-souring-treaties-protect-foreign-investors-arbitration.

Economist, The. 2014b. "The Dozy Watchdogs," December 13. http://www.economist.com/news/briefing/21635978-some-13-years-after-enron-auditors-still-cant-stop-managers-cooking-books-time-some.

Economist, The. 2015. "The Great Distortion," May 16. http://www.economist.com/news/leaders/21651213-subsidies-make-borrowing-irresistible-need-be-phased-out-great-distortion.

Economist, The. 2016. "The Biggest Loophole of All, " February 20. http://www.economist.com/news/international/21693219-having-launched-and-led-battle-against-offshore-tax-evasion-america-now-part.

Egan, Mark, Gregor Matvos, and Amit Seru. 2017. "The Market for Financial Adviser Misconduct." NBER Working Paper 22050.

Eisinger, Jesse. 2017. *The Chickenshit Club: The Justice Department and the Failure to Prosecute White Collar Criminals.* New York, NY: Simon and Schuster.

Felcher, Marla. 2001. *It's No Accident: How Corporations Sell Dangerous Baby Products.* Monroe, ME: Common Courage Press.

Freifeld, Karen. 2017. "Misconduct Rife in Derivatives—Ex-CFTC Enforcement Chief." *Reuters*, March 24. http://www.reuters.com/article/us-cftc-enforcement-goelman-idUSKBN16V1D0.

Friedman, Milton. 1970. "The Social Responsibility of Business Is to Increase Its Profits." *New York Times*, September 13.

Garrett, Brandon L. 2016. *Too Big to Jail: How Prosecutors Compromise with Corporations.* Cambridge, MA: Belknap Press.

Garrett, Brandon L., and Jon Ashley. 2017. Corporate Prosecution Registry. http://lib.law.virginia.edu/Garrett/corporate-prosecution-registry/about.html.

Gilson, Ronald J., and Jeffrey N. Gordon. 2013. "The Agency Costs of Agency Capitalism: Activist Investors and the Revaluation of Governance Rights." *Columbia Law Review* 113(4): 863–927.

Graham, John R., Campbell R. Harvey, and Shiva Rajgopal. 2005. "The Economic Implications of Corporate Financial Reporting." *Journal of Accounting and Economics* 40(1–3): 3–73.

Hansmann, Henry, and Reinier Kraakman. 2001. "The End of History for Corporate Law." *Georgetown Law Journal* 89(2): 439–68.

Hirshleifer, David, and Siew Hong Teoh. 2009. "Systemic Risk, Coordination Failures, and Preparedness Externalities: Applications to Tax and Accounting Policy." *Journal of Financial Economic Policy* 1(2): 128–42.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7: 24–52.

Independent Directors of the Board of Wells Fargo & Company. 2017. *Sales Practices Investigation Report.* New York: Independent Directors of the Board of Wells Fargo & Company.

Johnson, Simon, and James Kwak. 2010. *13 Bankers: The Wall Street Takeover and the Next Financial Meltdown*. New York, NY: Pantheon.

Krippner, Greta R. 2011. *Capitalizing on Crisis.* Cambridge, MA: Harvard University Press.

Leising, Matthew. 2017. "Hunting for Dirty Deeds in the $34 Trillion U.S. Futures Market." *Bloomberg Businessweek*, February 16. https://www.bloomberg.com/news/articles/2017-02-16/hunting-for-dirty-deeds-in-the-34-trillion-u-s-futures-market.

Lessig, Lawrence. 2012. *Republic, Lost: How Money Corrupts Congress—And a Plan to Stop It.* New York, NY: Twelve.

Lipton, Eric, and Ben Protess. 2013. "Banks' Lobbyists Help in Drafting Financial Bills." *New York Times*, May 23. https://dealbook.nytimes.com/2013/05/23/banks-lobbyists-help-in-drafting-financial-bills/?mcubz=0&_r=0.

Lucca, David, Amit Seru, and Francesco Trebbi. 2014. "The Revolving Door and Worker Flows in Banking Regulation." *Journal of Monetary Economics* 65: 17–32.

Manne, Henry G. 1965. "Mergers and the Market for Corporate Control." *Journal of Political Economy* 73(2): 110–20.

Mayer, Colin. 2013. *Firm Commitment: Why the Corporation Is Failing Us and How to Restore Trust in It.* Oxford University Press.

McCarty, Nolan. 2013. "Complexity, Capacity, and Capture." In *Preventing Regulatory Capture: Special Interest Influence and How to Limit It*, edited by Daniel Carpenter and David Moss, 99–123. Cambridge University Press.

Mian, Atif, and Amir Sufi. 2015. *House of Debt: How They (and You) Caused the Great Recession, and How We Can Prevent It from Happening Again.* University of Chicago Press.

Narayanan, M. P. 1985. "Managerial Incentives for Short-Term Results." *Journal of Finance* 40(5): 1469–84.

Norris, Floyd. 2013. "Masked by Gibberish, the Risks Run Amok." *New York Times,* March 21. http://www.nytimes.com/2013/03/22/business/behind-the-derivatives-gibberish-risks-run-amok.

html?mcubz=0.

OECD. 2015. *Measuring and Monitoring BEPS, Action 11: 2015 Final Report*. OECD/G20 Base Erosion and Profit Shifting Project series. Paris, France: OECD Publishing.

Pagano, Marco, and Paolo F. Volpin. 2005. "The Political Economy of Corporate Governance." *American Economic Review* 95(4): 1005–30.

Partnoy, Frank. 2009. *Infectious Greed: How Deceit and Risk Corrupted the Financial Markets*. New York, NY: PublicAffairs.

Partnoy, Frank. 2016. "What's (Still) Wrong with Credit Rating Agencies." San Diego Legal Studies Paper 17–285.

Partnoy, Frank, and Jesse Eisinger. 2013. "What's Inside America's Big Banks." *Atlantic*, January/February. https://www.theatlantic.com/magazine/archive/2013/01/whats-inside-americas-banks/309196/.

Piskorski, Tomasz, Amit Seru, and James Witkin. 2015. "Asset Quality Misrepresentation by Financial Intermediaries: Evidence from the RMBS Market." *Journal of Finance* 70(6): 2635–78.

Ramanna, Karthik. 2015. *Political Standards: Corporate Interest, Ideology, and Leadership in the Shaping of Accounting Rules for the Market Economy*. University of Chicago Press.

Reda, James F., David M. Schmidt, and Kimberly A. Glass. 2016. "Study of 2015 Short- and Long-Term Incentive Design Criteria Among Top 200 S&P 500 Companies." Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2916206.

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

Ronen, Joshua. 2010. "Corporate Audits and How to Fix Them." *Journal of Economic Perspectives* 24(2): 189–210.

Sawyer, Kim R., Jackie Johnson, and Mark Holub. 2010. "The Necessary Illegitimacy of the Whistleblower." *Business and Professional Ethics Journal* 29(1–4): 85–107.

Skeel, David A. Jr., and Thomas H. Jackson. 2012. "Transaction Consistency and the New Finance in Bankruptcy." *Columbia Law Review* 112(1): 152–202.

Stein, Jeremy C. 1989. "Efficient Capital Markets, Inefficient Firms: A Model of Myopic Corporate Behavior." *Quarterly Journal of Economics* 104(4): 655–69.

Stigler, George J. 1971. "The Theory of Economic Regulation." *Bell Journal of Economics and Management Science* 2(1): 3–21.

Story, Louise, and Stephanie Soul. 2015. "Towers of Secrecy: Piercing the Shell Companies." *New York Times*, February 7–14. https://www.nytimes.com/news-event/shell-company-towers-of-secrecy-real-estate?mcubz=0.

Taub, Jennifer. 2009. "Able But Not Willing: The Failure of Mutual Fund Advisers to Advocate for Shareholders' Rights." *Journal of Corporation Law* 34(3): 843–93.

Teoh, Siew Hong, Ivo Welch, and T. J. Wong. 1998. "Earnings Management and the Underperformance of Seasoned Equity Offerings." *Journal of Financial Economics* 50(1): 63–99.

Wayne, Leslie. 2012. "How Delaware Thrives as a Corporate Tax Haven." *New York Times*, June 20. http://www.nytimes.com/2012/07/01/business/how-delaware-thrives-as-a-corporate-tax-haven.html?mcubz=0.

Zingales, Luigi. 2015. "Presidential Address: Does Finance Benefit Society?" *Journal of Finance* 70(4): 1327–63.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 151–174

The Causes and Costs of Misallocation

Diego Restuccia and Richard Rogerson

Why do living standards differ so much across countries? This is one of the long-standing questions in economics. A consensus in the development literature is that differences in productivity are a large, if not necessarily the dominant, source of these differences: that is, even after adjusting for differences in the quantity and quality of factors of production such as capital and labor, poor countries produce much less output per worker than rich countries, and this difference accounts for much of the variation in income per capita across countries.[1] But what accounts for productivity differences across countries? One explanation is that frontier technologies and best practice methods are slow to diffuse to low-income countries. The recent literature on misallocation, which is the focus of this article, offers a distinct but complementary explanation: low-income countries are not as effective in allocating their factors of production to their most efficient use.

Casual empiricism suggests that both slow diffusion and misallocation are potentially relevant. A visit to any less-developed country reveals that much production, whether in agriculture, manufacturing, or services, seems to use outdated

[1] Early contributions making this point include Klenow and Rodriguez-Clare (1997), Prescott (1998), and Hall and Jones (1999). See also the surveys of Caselli (2005) and Jones (2016).

■ *Diego Restuccia is a Professor of Economics and a Canada Research Chair, University of Toronto, Toronto, Canada, and a Research Associate at the National Bureau of Economic Research, Cambridge, Massachusetts. Richard Rogerson is the Charles and Marie Robertson Professor of Public and International Affairs, Woodrow Wilson School, Princeton University, Princeton, New Jersey, and a Research Associate at the National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are diego.restuccia@utoronto.ca and rdr@princeton.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.151

doi=10.1257/jep.31.3.151

methods. But many studies and anecdotes detail how corruption, regulation, or direct government involvement distort the allocation of resources from their most efficient use, especially in poorer economies. More generally, the notion that the allocation of inputs across establishments is an important component of aggregate productivity is reinforced by studies in the United States and elsewhere that find reallocation of inputs from less- to more-productive establishments to be an important component of aggregate productivity growth (for example, see Baily, Hulten, and Campbell 1992; Foster, Haltiwanger, and Syverson 2008).

Three key questions arise: First, how important is misallocation as a source of aggregate productivity differences across countries? Second, what are the main causes of misallocation? Third, beyond the direct cost of lower contemporaneous output, are there additional costs associated with misallocation? In this article, we provide our perspective on these three questions. It is not our intention to survey the literature, and as a result, we inevitably neglect many important references and contributions. We instead refer the reader to available survey articles of this literature, for instance Restuccia and Rogerson (2013) and Hopenhayn (2014).

Potential Sources of Misallocation

The nature of misallocation on which we focus is quite specific. Economists routinely study distortions that affect resource allocations along many dimensions, but we are specifically interested in distortions that affect the allocation of inputs *across* producers of a given good. For example, in the context of the standard neoclassical growth model, a proportional tax on income will distort household decisions regarding consumption and labor supply, and hence may be described as causing misallocation along these margins. But this type of misallocation, affecting the amounts of capital and labor used in production, is not the sort of misallocation we emphasize. Instead, we are interested in situations in which the allocation of a given amount of capital and labor across heterogeneous producers is distorted. This would happen, for example, when different producers of the same good are taxed at different rates.

An example will serve to fix ideas and facilitate exposition. Aggregate output is produced by many heterogeneous producers that differ in their individual levels of productivity.[2] Specifically, assume there are N potential producers of a homogeneous good and that producer i has a production function $y_i = A_i \cdot f(h_i, k_i)$, where y_i, is output, h_i is labor input, k_i is capital input, f is a strictly increasing and strictly concave production function, and variation in A_i reflects differences in productivity across producers. Assume also that there is a fixed cost for any producer who operates, measured in units of output and denoted by c. Given an aggregate amount of labor and capital, denoted by H and K respectively, there is a unique choice of which

[2] As summarized in Syverson (2011), large dispersion of productivity even within narrowly defined industries is a robust feature of reality.

producers should operate and how labor and capital should be allocated across them in order to maximize total output net of fixed operating costs.

Three conceptually distinct channels will affect the amount of output, and hence the overall level of productivity. The first channel, which we call the *technology* channel, reflects the values of the producer-level productivity A_i; if all of the A_i are larger, output will be greater. The second channel, which we call the *selection* channel, reflects the choice of which producers should operate. The third channel is the *misallocation* channel and reflects the choice of how capital and labor are allocated among those producers that operate. Conceptually, selection effects are a special case of misallocation, but from an empirical perspective we do not observe potential producers who do not operate, making it more difficult to measure selection effects without additional structure. An important theme in our discussion is that these three channels are not independent: any policy or institution that distorts the allocation of resources across producers—creating misallocation—will potentially generate additional effects through both the selection and technology channels.

In our example, output maximizing choices have the following form: a threshold rule determines which producers operate (that is, producers operate if the productivity level $A_i > \overline{A}$) and conditional upon operation, producers with higher values of A_i will be allocated a greater amount of labor and capital. The efficient allocation will induce a distribution of producer sizes. More specifically, the allocation of inputs that maximizes output will equate the marginal products of labor and capital across all producers with positive inputs. Thus, thinking about factors that interfere with equalization of marginal products is a useful way to identify possible sources of misallocation.

Many articles, spanning the fields of development economics, industrial organization, labor economics, finance, international economics, and others have documented specific sources of misallocation in particular contexts.[3] They serve to impress upon us the pervasiveness of misallocation. Rather than provide a laundry list of very specific potential sources of misallocation, we instead emphasize three general categories of factors.

First, misallocation may reflect statutory provisions, including features of the tax code and regulations. Specific examples would include provisions of the tax code that vary with firm characteristics (such as the size or age of the firm), tariffs applied to narrowly defined categories of goods, labor market regulations such as

[3]A list of studies on misallocation in specific areas could be extremely long, but we highlight a few examples. In the development literature, Banerjee and Duflo (2005) document credit market imperfections among manufacturers in India. De Mel, McKenzie, and Woodruff (2008) establish wedges between the marginal product of capital and borrowing rates among small producers in Sri Lanka using experimental methods. Besley and Ghatak (2010) survey work on property rights and misallocation. In industrial organization, Olley and Pakes (1996) study regulation in the US telecommunications industry and find an important role for misallocation. Caballero et al. (2008) document "zombie lending" practices in Japan, a process by which banks continue to extend credit to poorly performing businesses in order to avoid writing down bad loans. Heckman and Pagés (2004) summarize work on the effects of labor market regulations using microdata from Latin America and the Caribbean. Guner, Ventura, and Xu (2008) document the effects of product market regulation in Japan and India. Melitz and Redding (2014) summarize the literature on trade barriers and misallocation.

employment protection measures, product market regulations that restrict size or limit market access, and land regulations. Even a regulation that applies uniformly to all firms within an industry may generate misallocation within the industry. For example, a given employment protection measure will differentially affect expanding and contracting firms.

Second, misallocation may reflect discretionary provisions made by the government or other entities (such as banks) that favor or penalize specific firms. Such provisions are often referred to as "crony capitalism" or even "government corruption." Examples are subsidies, tax breaks, or low interest rate loans granted to specific firms, along with unfair bidding practices for government contracts, preferential market access, or selective enforcement of taxes and regulations.

Third, misallocation may reflect market imperfections. Examples include monopoly power, market frictions, and enforcement of property rights. A producer with monopoly power may produce less than the efficient level but charge a higher markup. A highly productive firm with little collateral may not be able to access enough capital to produce at the efficient level. Bloom et al. (2013) suggests that the size of highly productive firms in India is restricted by the inability to delegate management outside of the family on account of poor enforcement of property rights. Lack of land titling may affect the allocation of land.

There are three messages that we want the reader to take away from this overview. First, the set of plausible underlying sources of misallocation is wide-ranging. Second, many sources are very narrow in scope—being particular to specific sectors, types of firms, or even regions. And third, many of these sources, especially those reflecting discretionary provisions, are not amenable to systematic measurement. This combination makes life challenging for any researcher interested in assessing the aggregate importance of misallocation.

Measuring Misallocation: Methodology

Misallocation seems pervasive. But is it quantitatively important? To address the question of whether misallocation is an important source of cross-country differences in total factor productivity, the literature has adopted two main approaches, which we label the *direct* and the *indirect* approaches.

The essence of the direct approach is to focus on specific sources of misallocation and to assess their consequences. One source of information is quasi-natural experiments that shed light on a particular source of misallocation. While some studies have successfully followed this path, as a practical matter, the scope for this type of assessment seems to be somewhat limited. As a result, the typical study employing the direct approach seeks to measure the source of misallocation and assess its quantitative effects via a structural model. This approach has a long tradition in public finance as a way to measure the distortions from various taxes. Of course, a researcher must be aware that details of the structural model may have important effects on the findings. However, we stress that evaluating the extent of misallocation necessarily requires computing a counterfactual—how much

additional output could be generated by reallocating inputs among producers. One cannot entirely avoid structure in answering this question.

But the direct approach faces another challenge. Implementing it requires quantitative measures of the underlying source of misallocation. If statutory provisions are the key source of misallocation, then this is perhaps not a problem. However, if the most important sources of misallocation reflect discretionary provisions, then measurement may be very difficult. Even if regulation is an important source of misallocation in aggregate, the highly specialized and complex nature of regulation within specific industries may still make it very difficult to develop and analyze an appropriate structural model.

In contrast, the indirect approach seeks to identify the extent of misallocation without identifying the underlying source of the misallocation. As noted earlier in our simple example, the efficient allocation of inputs equates marginal products across all active producers. Thus, directly examining variation in marginal products provides the opportunity to measure the amount of misallocation without specifying the underlying source of misallocation. This approach also requires some structure, but unlike the direct approach it does not require specifying a full model. In our simple example, given cross-section data on output, labor, and capital, it is sufficient to specify the production function f in order to directly compute the implied amount of misallocation. To see why, note that with data on y, k, and h for each producer and a production function f, we can infer the A_i. Given a production function f and the A_i, we can directly solve for the allocation of inputs among producers that would maximize output. Comparing this to actual output provides an assessment of the extent of misallocation. Note that because this exercise takes the set of producers as given, it does not address selection. So even though selection effects are conceptually akin to what we have called misallocation, this procedure will only isolate the misallocation effect.

Although the indirect approach requires less structure than the direct approach, it faces one key challenge. In more general frameworks, efficient allocations need not entail equality of marginal products across producers at every point in time. If inputs are chosen before the realization of producer-specific shocks, or if there are adjustment costs, then this condition need not hold. Also, measurement error in firm-level data will lead us to infer variation in marginal products across producers even when none truly exists. We later discuss these issues in more detail.

How Important is Misallocation? Results Using the Indirect Approach

In Restuccia and Rogerson (2008), we used a version of the Hopenhayn (1992) industry equilibrium model calibrated to match features of the US economy to explore the extent to which misallocation caused by firm-specific taxes and subsidies would impact aggregate total factor productivity. These firm-specific taxes and subsidies were hypothetical, but chosen as a representation of the many different factors that might generate misallocation. In one of our scenarios, termed "correlated distortions," high-productivity establishments are systematically taxed and low-productivity establishments are systematically subsidized. We showed that

this can substantially depress total factor productivity. One key message from this research is that for misallocation to have large effects, it needs to depress inputs systematically at high-productivity producers. It follows that studies identifying misallocation among relatively small and less-productive enterprises may not be particularly relevant in terms of assessing aggregate effects.

The Indirect Approach

Whereas in Restuccia and Rogerson (2008) we analyzed misallocation from hypothetical policy distortions, Hsieh and Klenow (2009) noted that the extent of misallocation could be estimated given appropriate microdata and some structure. Their procedure essentially follows the strategy described in the previous section but in a setting where each firm produces a distinct variety of goods that are valued by consumers according to a constant elasticity of substitution aggregator. Each producer behaves as a monopolistic competitor when deciding its level of output, but markets for labor and capital are competitive. The implied demand structure is important because it allows the authors to infer total factor productivity when the data includes only total revenue (as opposed to physical output).

When Hsieh and Klenow (2009) apply their method to four-digit manufacturing industries in China, India, and the United States, they find large effects of misallocation on total factor productivity. In particular, if misallocation were eliminated, total factor productivity in manufacturing would increase by 86–110 percent in China, 100–128 percent in India, and 30–43 percent in the United States. Taken at face value, these results indicate that misallocation is quantitatively important, even in a high-income economy like the United States, and that it is an important factor in accounting for productivity differences across rich and poor countries. These estimates are for the manufacturing sector, not the overall economy. Available evidence suggests that cross-country differences in manufacturing productivity tend to be much smaller than aggregate productivity differences. Hsieh and Klenow (2009) estimated that total factor productivity differences in manufacturing between the United States and China and India during the relevant period are on the order of 130 and 160 percent respectively, in contrast to total factor productivity differences on the order of 300 and 600 percent at the aggregate level.

We note that these productivity losses from misallocation assume that all dispersion in revenue marginal products across producers within a sector is the result of distortions or institutions that can be acted upon by policy. To the extent that some differences need not reflect misallocation due to policies, their estimates overstate the total amount of misallocation. We return to this issue later.

Although the Hsieh and Klenow (2009) approach measures misallocation without identifying the source of the misallocation, their analysis does nonetheless allow them to examine how the extent of misallocation is correlated with various observables. For example, state ownership in China is intimately related with misallocation, in that state-owned firms are much larger than efficiency would dictate. Another important finding is that high-productivity producers are too small in all three economies, but the size of this effect is stronger in China and India than in the United States. Bento and Restuccia (forthcoming) corroborate this finding for a

larger set of developing countries: the extent to which more-productive plants face greater implicit taxes is strongly related to GDP per capita across countries.

Limitations of the Indirect Approach

The indirect approach essentially assumes a production structure and then uses the data to estimate wedges in the first-order conditions that characterize an efficient allocation. This approach interprets the wedges as reflecting distortions to efficient allocations. But related to our earlier discussion, there are good reasons to be wary of this interpretation. We discuss three specific reasons that we believe are potentially significant. We note that Hsieh and Klenow (2009) acknowledged and attempted to address each of them.

The first issue concerns the nature of heterogeneity in production functions across producers. With enough freedom to choose heterogeneous production functions across producers, data on inputs and outputs would not allow one to infer differences in marginal products. But what about some restricted and seemingly reasonable degrees of heterogeneity? For example, the benchmark results in Hsieh and Klenow (2009) assume all producers within a sector use the same Cobb–Douglas production function. It follows that capital-to-labor ratios are equal for all producers in an efficient allocation, implying that any variation in capital-to-labor ratios will be interpreted as misallocation. An alternative interpretation is that producers use different production methods so that capital shares in the Cobb–Douglas production function are heterogeneous across producers. In the extreme, all differences in capital-to-labor ratios reflect heterogeneity in producer-level production functions, rather than misallocation. Hsieh and Klenow (2009) show that although this alternative interpretation implies less misallocation, the remaining misallocation still implies large productivity losses. This result implies that the dominant sources of distortions act symmetrically on labor and capital so that the capital to labor ratio is roughly unaffected by distortions.

The second issue we consider is adjustment costs. A voluminous literature estimates substantial adjustment costs for both labor and capital at the individual producer level (for example, see Cooper and Haltiwanger 2006; Bloom 2009; and the survey in Bond and Van Reenen 2007). This raises the possibility that marginal products of capital and labor in production differ across producers because of adjustment costs and transitory firm-specific shocks. Being mindful of this issue, Hsieh and Klenow's (2009) preferred interpretation of their findings is to focus on the *differences* in misallocation across economies, rather than the levels per se. The idea is that some amount of "base level" misallocation is appropriately understood as the result of adjustment costs or some other misspecification, and that a reasonable starting point is to assume that this level is the same across economies. This moderates their estimates of the amount of misallocation: if China and India were to reduce misallocation to the level found in the United States, total factor productivity in manufacturing in those countries would increase by 31–51 percent and 40–59 percent, respectively. While smaller than the earlier values, it remains true that misallocation can account for almost half of the observed total factor productivity differences in manufacturing.

But is it reasonable to argue that all economies have some common level of measured misallocation that should be ignored in this context? Asker, Collard-Wexler, and De Loecker (2014) argue that the answer to this question is no. They show that observed differences in the dispersion of marginal revenue products can be consistent with efficient allocations if there are adjustment costs on capital coupled with transitory firm-level shocks that are more variable in poorer countries. While we believe that this study serves as an important cautionary note regarding the indirect approach, two remarks are important. First, it is necessary to ask why idiosyncratic shocks are more variable in poorer countries—if the higher variability of shocks reflects greater variability in the policy environment then it seems appropriate to interpret the higher dispersion of marginal revenue products in poorer countries as reflecting misallocation. Second, it highlights the need to examine misallocation using panel data at the establishment level, instead of cross-section data. If measured misallocation is due to adjustment costs, it will generate specific time-series patterns. More generally, with panel data, researchers could carry out the indirect approach on specifications that explicitly include adjustment costs. David and Venkateswaran (2017) carry out exactly this type of analysis using panel data from China under the assumption that capital adjustment costs are convex. While adjustment costs and idiosyncratic policy distortions can both generate the cross-sectional dispersion in the marginal product of capital across firms, they have opposing effects on the autocorrelation of investment. Using dynamic moments from their panel dataset, the authors show that most of the cross-sectional variation in marginal revenue products is due to policy distortions with a relatively minor share due to adjustment costs. This result appears robust to considering nonconvex adjustment costs because at the annual frequency, inaction due to fixed costs is estimated to be minor. But more analysis of this type using panel data is an important priority for future research.

Third, the higher dispersion of marginal products in China and India could reflect greater amounts of measurement error in these countries relative to the United States. Hsieh and Klenow (2009) carry out several calculations to assess this possibility, which, while not conclusive, do not support such an interpretation. Recent work by Bils, Klenow, and Ruane (2017) goes much further. They use the panel component of the datasets for India and the United States used in Hsieh and Klenow (2009) to estimate measurement error in each country and infer the extent of differences in productivity due to misallocation after accounting for measurement error. They have three main findings. First, measurement error accounts for a substantial amount of the dispersion in marginal revenue products. Second, the contribution of measurement error is becoming more important over time in the United States but is relatively stable in India. And third, after accounting for measurement error, the contribution of misallocation to understanding productivity differences between India and the United States is very similar to what Hsieh and Klenow (2009) found in their original analysis, that is manufacturing total factor productivity gains of 40–60 percent in India relative to the United States.

While progress is being made in extending the indirect method to address the limitations discussed, we also think it is useful to develop alternative approaches. For example, Bartlesman, Haltiwanger, and Scarpetta (2013) focus on the covariance

between firm size and productivity, and how it is affected by firm-specific taxes and subsidies. They assume a specification that implies cross-sectional differences in marginal products even in an efficient allocation, and calibrate their model so that moments of the US cross-sectional data on revenue productivity dispersion and employment are consistent with efficiency. They use the calibrated model to assess the amount of misallocation in manufacturing in a sample of seven other economies—the United Kingdom, France, Germany, Netherlands, Romania, Hungary, and Slovenia—during the 1990s. Rather than inferring the actual distortions faced by each firm in their dataset, they infer a statistical representation of distortions that matches salient moments of the data. Relative to the United States, they find that the effect of misallocation on total factor productivity ranges from 3 percent in Germany to 12 percent in Romania. Their limited choice of countries was dictated by the desire to have data that was consistently collected across countries, so drawing broad conclusions about difference across countries is not possible. But studies like this open the possibility of comparing the estimates of misallocation for a given country based on different methods.

Further Indirect Evidence on Misallocation in Different Countries and Sectors

The analysis in Hsieh and Klenow (2009) found important effects of misallocation within manufacturing in China and India relative to the United States. A variety of studies have extended this finding to other countries and other sectors. Busso, Madrigal, and Pagés (2013) carry out a comparable analysis of manufacturing in ten Latin American countries and conclude that differences in misallocation between these economies and the United States is an important source of total factor productivity gaps in manufacturing. Kalemli-Ozcan and Sørensen (2016) study misallocation of capital among private manufacturing firms in 10 African countries. Their sample also includes firms from India, Ireland, Spain, and South Korea that can be used as benchmarks. Subject to the caveat of small sample sizes, they find that capital misallocation in Africa is significantly higher than in developed countries, though not as severe as in India.

The above results all pertain to the manufacturing sector. Relatively few papers have addressed misallocation in the service sector. Busso, Madrigal, and Pagés (2013) include analyses of specific service sectors, such as retail, and find that misallocation in services sectors is much larger than in manufacturing. De Vries (2014) finds very large misallocation in the retail sector in Brazil. Dias, Marques, and Richmond (2016a) study misallocation in manufacturing and services in Portugal and also find that misallocation is much larger in services. One limitation of these studies is that they do not include a benchmark, such as the US economy. If misallocation measures for the US economy are also larger in service sectors than in manufacturing, then it is not clear if misallocation differences are indeed more severe in service sectors. Also, an important caveat is that output in a number of relevant service sectors, such as education, health care, and banking, is likely to be very poorly measured.

The agricultural sector is of particular importance in comparing the world's richest and poorest economies as this is where productivity gaps are greatest and a

large share of labor in poor countries is allocated to agriculture (Gollin et al. 2002; Restuccia, Yang, and Zhu 2008). Caselli (2005) reports that differences in output per worker, expressed in terms of the ratio of countries in the 90th percentile to the 10th percentile of the income distribution, were 22 at the aggregate level, 4 in nonagriculture, and 45 in agriculture.

Adamopoulos and Restuccia (2014) document a long list of policies and institutions in the agricultural sector in developing countries that can potentially create misallocation. They also document striking differences in the distribution of farm sizes across countries with the typical operational land scale of a farm in poor countries being only 2 to 3 percent of the operational size in rich countries. The authors develop a model of agriculture and nonagriculture extended to produce a nondegenerate endogenous distribution of farms sizes in agriculture and consider abstract representations of distortions to match the observed distribution of farm sizes across countries. They find that the misallocation created by farm-size distortions can account for much of the farm-size and productivity differences in agriculture between rich and poor countries. Additionally, they show that the implied farm-size distortions are consistent with data on within-country variation in crop-specific price distortions and their correlation with farm size.

Restuccia and Santaeulalia-Llopis (2017) study misallocation across household farms in Malawi. They have data on the physical quantity of outputs and inputs as well as measures of transitory shocks and so are able to measure farm-level total factor productivity. They find that the allocation of inputs is relatively constant across farms despite large differences in measured total factor productivity, suggesting a large amount of misallocation. In fact, they found that aggregate agricultural output would increase by a remarkable factor of 3.6 if inputs were allocated efficiently. Their analysis also suggests that institutional factors that affect land allocation are likely playing a key role. Specifically, they compare misallocation within groups of farmers that are differentially influenced by restrictive land markets. Whereas most farmers in Malawi operate a given allocation of land, other farmers have access to marketed land (in most cases through informal rentals). Using this source of variation, Restuccia and Santaeulalia-Llopis find that misallocation is much larger for the group of farmers without access to marketed land: specifically, the potential output gains from removing misallocation are 2.6 times larger in this group relative to the gains for the group of farms with marketed land.

Other studies also document misallocation in agriculture. For instance, Adamopoulos, Brandt, Leight, and Restuccia (2017) study the case of China between 1993 and 2002, where the land market is severely restricted by the "household responsibility system." Land ownership and allocation decisions reside with the collective village, and use rights of land are distributed uniformly among household members registered in the village. While there are no explicit restrictions on land rental in China, fear of redistribution leads to implicit "use it or lose it" rules. In this context, farm operational scales are essentially limited to the use rights of land for each household, and hence, not surprisingly, the authors find that land allocations are unrelated to farm productivity. In particular, eliminating misallocation in this context is found to increase agricultural productivity by 1.84-fold, with 60 percent of

this gain arising from reallocation of factors across farms within villages. Exploiting the panel dimension of the data to remove potential transitory variation in farm productivity, the authors show that reallocation gains are still substantial, representing 81–86 percent of the cross-sectional productivity gains.

Chen, Restuccia, and Santaeulàlia-Llopis (2017) study the case of Ethiopia, where the current land market institutions are the result of a long history of divisive land relationships and conflicts. Land ownership resides with the state, and local authorities allocate land-use rights equally among households, controlling for soil quality and household size. Using detailed micro household-level data, the authors document substantial misallocation of land and other factors of production in the agricultural sector. An efficient reallocation of inputs can increase aggregate agricultural productivity by a factor of 2.4, with 75 percent of this increase derived from reallocation within zones (counties) in Ethiopia. The authors also exploit regional variation in the extent of rented land due to differential implementation of a land certification program that started in the early 2000s. Even though most rentals still occur between family members and relatives, they found that regions with more land rentals have significantly less misallocation: a 1 percentage point higher share of land rental is associated with a 0.8 percentage point decrease in the efficiency gain from reallocation.

Misallocation over Time

The results described so far have focused on differences in misallocation across countries at a point in time. It is also of interest to ask whether changes in misallocation over time within a country are an important source of changes in productivity over time. This is akin to connecting misallocation with growth accounting.

The literature has identified changes in misallocation as an important component of low-frequency movements in productivity in three contexts. Chen and Irrazabal (2015) show that misallocation decreased during Chile's decade-long period of growth following the crisis of the early 1980s, and was an important part of productivity growth during this time. Fujii and Nozawa (2013) show that capital misallocation in manufacturing became more pronounced after 1990 in Japan, a period characterized by poor productivity growth. And Gopinath, Kalemli-Ozcan, Karabarbounis, and Villegas-Sanchez (2015) find increased capital misallocation and roughly constant labor misallocation in Southern European countries subsequent to these countries joining the euro in 1999, a period of slower productivity growth in these countries. Note that changes in total factor productivity over time tend to be much smaller than differences in the cross-section, so even modest changes in misallocation can play a dominant role in the context of the time series changes observed in the data.[4]

A promising avenue for further study is to focus on changes in misallocation during periods in which important policy or regulatory changes occurred that

[4] See also Reis (2013) and Dias, Marques, and Richmond (2016b) for the case of Portugal, and Calligaris (2015) for Italy. Ziebarth (2013) is an interesting analysis of long-run changes in the context of the United States. In particular, he found that misallocation levels among US manufacturers in the late 19th century were similar to those in present-day India and China.

one might reasonably believe have important effects on misallocation. Hsieh and Klenow (2009) took a first step in this direction. They found a decrease in misallocation in China during the period of 1998 to 2005, a finding consistent with the view that various reforms enacted during this period served to lessen the importance of distortions. Interestingly, despite widespread reform in other sectors, land market institutions have remained essentially the same in China, and Adamopoulos et al. (2017) found that misallocation in the agricultural sector in China has remained roughly constant for the period of study (1993–2002).

Hsieh and Klenow (2009) found that misallocation in India worsened over the period from 1987 to 1994, a result which seems puzzling given the nature of reforms enacted there. One important reform during this time was the elimination of the license "raj" system, a system of controls on the entry of firms into the manufacturing sector. Bollard, Klenow, and Sharma (2013) pursued this further and found that although this period witnessed rapid productivity growth for their sample of very large firms, little of the productivity growth was due to changes in misallocation. There are of course multiple interpretations of this finding; perhaps the raj system was not an important source of misallocation among large firms, or perhaps it is not even an important source of misallocation overall. Alternatively, as noted earlier, the indirect method might not be isolating true misallocation. A recurring theme in this work is the need to reconcile results based on differing approaches.

The research by Bartleseman, Haltiwanger, and Scarpetta (2013) described earlier also included a time series component. They found that misallocation decreased over the period of the 1990s in the transition economies of Eastern Europe. This finding is also consistent with the notion that increased market reforms were leading to less misallocation, but the extent of the change is somewhat modest, increasing productivity by a few percentage points.

Several papers have assessed changes in misallocation over the business cycle, typically focusing on fairly dramatic episodes such as crises or protracted recessions. Oberfield (2013) studies misallocation in Chile during the crisis of the early 1980s, Sandleris and Wright (2014) examine misallocation in Argentina during its crisis in the early 2000s, and Ziebarth (2015) assessed misallocation during the US Great Depression. All of these authors find that misallocation increased sharply in each of these episodes and accounted for a large part of measured drops in aggregate total factor productivity. However, in our view, changes in misallocation measures at business cycle frequency need to be treated with extreme caution. As emphasized earlier, these measures can be heavily influenced by adjustment costs that may give rise to factor hoarding. To us, it remains very much an open question whether true misallocation of resources increases during these episodes.

Causes of Misallocation: The Direct Approach

The broad message that emerges from the many studies that employ the indirect approach is that misallocation is an important source of productivity differences across countries. But what is the underlying source of this misallocation? To

answer this question, we discuss the efforts to isolate causes of misallocation using the direct approach. Our goal is to assess the aggregate importance of misallocation attributed to several categories of distortions, particularly with an eye toward asking whether we can isolate factors that might generate effects of the magnitude found using the indirect method. In this regard, the current state of this literature is somewhat disappointing. The existing literature has identified some factors that can account for large effects of misallocation in agriculture. But it has yet to identify any particular factor that can account for the magnitudes of misallocation found in manufacturing.

Regulation

One of the earliest studies of misallocation due to regulation is the analysis of firing costs in Hopenhayn and Rogerson (1993). Firing costs are an adjustment cost created by policy, and the resulting variation in marginal products therefore reflects true misallocation. Using a quantitative version of the model in Hopenhayn (1992), these authors find that firing costs equal to one year's wages will lead to steady-state productivity losses of roughly 2 percent.[5] While these effects are comparable to a year of productivity growth for a typical country, they are nonetheless small relative to the magnitude of cross-country differences that we offered as the key motivating observation for the misallocation literature.

A potentially broader category of policies, what Guner, Ventura, and Xu (2008) call "size-dependent policies," reflects measures that implicitly levy higher taxes on firms that are larger in terms of sales, labor, or capital. Examples include regulations that only become effective beyond some employment threshold, outright restrictions on the number of employees, or restrictions on the amount of physical space that a retail establishment may operate. They analyze simple but abstract versions of such policies, and find that while they can have large effects on the number of firms and the firm size distribution, they have relatively small effects on total factor productivity.[6]

A large literature in development economics has studied duality and informality as a source of low productivity in poor countries (Lewis 1954; Rauch 1991; La Porta and Shleifer 2014). This literature is a natural predecessor to quantitative studies of misallocation, as one of its key ideas is that development requires the reallocation of resources out of subsistence and informal activities into "modern" activities. Busso, Fazzio, and Levy (2012) study the relation between productivity and informality in Mexico using detailed microdata. They exploit a precise

[5]Lagos (2006) uses a Mortensen–Pissarides matching model to study how labor market policies such as unemployment insurance and employment protection affect productivity via selection effects. He finds that changes in the replacement rate and firing costs decrease aggregate total factor productivity on the order of 2–3 percent.

[6]In related work, Garcia-Santana and Pijoan-Mas (2014) study the quantitative effect of small-scale reservation laws in India, a form of firm-size restriction. In a calibrated version of their model using plant-level data for India, eliminating these laws increases manufacturing output by almost 7 percent and manufacturing total factor productivity (TFP) by 2 percent. Also, Gourio and Roys (2014) and Garicano, Lelarge, and Van Reenen (2016) study the effects of size-dependent labor regulations using plant-level data from France where firms with 50 or more employees face substantial additional labor regulations.

definition of informality based on the institutions and laws that regulate relations between workers and firms, which in the case of Mexico involves the asymmetric regulation of salaried and nonsalaried workers, and separate notions of informality and illegality. Using these definitions, the authors document productivity, size, and misallocation distributions for each group. Controlling for firm size and legal status, informal firms are much less productive than formal firms, yet command a large share of resources and hence contribute significantly to low productivity in Mexico. While this study documents the correlation between informality and productivity, an important limitation is that it does not address causation. Related to this issue, Leal Ordóñez (2014) calibrates a model using data from Mexico that assumes firms can avoid regulation by choosing to hire capital below a certain threshold. His model accounts for the large share of activity in the informal sector but he finds that making enforcement uniform would only increase total factor productivity by slightly more than 4 percent (see also D'Erasmo and Moscoso Boedo 2012).

Government regulation can also hinder the reallocation of individuals across space. Hsieh and Moretti (2015) study misallocation of individuals across 220 US metropolitan areas from 1964 to 2009. They document a doubling in the dispersion of wages across US cities during the sample period. Using a model of spatial reallocation, they show that the increase in wage dispersion across US cities represents a misallocation that contributed to a loss in aggregate GDP per capita of 13.5 percent. They argue that across-city labor misallocation is directly related to housing regulations and the associated constraints on housing supply. Fajgelbaum, Morales, Suárez Serrato, and Zider (2015) study how the spatial allocation of workers and firms responds to US state taxes. They find that eliminating tax dispersion across US states produces modest increases in output, but note that this in part reflects the fact that dispersion in taxes across US states is not so large. Tombe and Zhu (2015) provide direct evidence on the frictions of labor (and goods) mobility across space and sectors in China and quantify the role of these internal frictions and their changes over time on aggregate productivity. The reduction of internal migration frictions is key and together with internal trade restrictions account for about half of the growth in China between 2000 and 2005.

Market activity can also be regulated via state-owned enterprises. The misallocation of resources in manufacturing between private and state-owned enterprises in China is a key source of productivity losses in the analysis of Song, Storesletten, and Zilibotti (2011). More recently, Brandt, Tombe, and Zhu (2013) study the importance of misallocation within the nonagricultural sector across state and nonstate enterprises and across provinces over time in China. They find that misallocation reduces nonagricultural total factor productivity by an average of 20 percent for the period 1985–2007. More than half of this productivity loss is due to within-province misallocation of capital between state and nonstate sectors. While across-province distortions remain fairly constant over time and there is a reduction in the share of state-owned enterprises over time, the authors find increased state/nonstate capital misallocation between 1998 and 2007. We are not aware of comparable studies for countries other than China.

Property Rights

A long tradition in development economics emphasizes property rights as a key institution shaping resource allocation and productivity (Besley and Ghatak 2010). Land reforms are common in developing countries (de Janvry 1981; Banerjee 1999; Deininger and Feder 2001) and represent an important example. They are often associated with a limit on farm size and restrictions on land markets so as to redistribute land from large landholders to landless and smallholder households. Adamopoulos and Restuccia (2015) study an example of such a comprehensive land reform in the Philippines using a quantitative model and panel microdata on farms that cover the period before and after the reform. They find that the reform substantially reduced farm size and agricultural productivity (reductions of 34 and 17 percent, respectively). The negative productivity effect reflects both a selection effect and a misallocation effect. Full enforcement of the farm size cap would have doubled the reduction in agricultural productivity.[7]

Trade and Competition

The effect of trade policy on aggregate productivity has been studied through the lens of models that extend the work of Eaton and Kortum (2002) and Melitz (2003). The key point is that tariffs and other forms of trade protection distort the allocation of resources across heterogeneous producers. Several studies provide model-based estimates of these effects, as surveyed in Kehoe, Pujolás, and Rossbach (forthcoming). An early example is Eaton, Kortum, and Kramarz (2011), who studied the effect of a 10 percent reduction in trade costs for all countries. Caliendo and Parro (2015) study the effects of NAFTA using this type of model. These studies find modest productivity effects.[8] But importantly, other studies have tackled the issue of trade liberalization and productivity directly by studying episodes of trade reform and viewing them as quasi-natural experiments. Two early examples are Pavcnik (2002) and Trefler (2004).[9] Pavcnik (2002) studies productivity changes in a micro-level panel dataset for Chile during an episode of substantial reductions in trade barriers that exposed plants to foreign competition. She isolates the contribution of trade to productivity growth by exploiting the variation in outcomes between plants in the import-competing/export-oriented sectors and plants in the nontraded sector. She finds that productivity increased by 19 percent and that roughly two-thirds of this was due to reallocation of resources from less- to more-productive producers. Trefler (2004) studies the Canada–United States Free Trade

[7]Similarly, de Janvry, Emerick, Gonzalez-Navarro, and Sadoulet (2015) study a land reform in Mexico in the 1990s in which farmers were given ownership certificates of land, removing the pre-existing link between land rights and land use, and show substantial labor and land reallocations associated with the reform.

[8]Waugh (2010) uses a version of the Eaton and Kortum (2002) model to infer trade barriers using data on observed trade flows and finds that eliminating trade restrictions substantially reduces cross-country income and productivity disparity. Tombe (2015) similarly argues that trade barriers are an important determinant of cross-country differences in productivity.

[9]Other examples include Bernard, Jensen, and Schott (2006) for the United States, Fernandes (2007) for Colombia, and Topalova and Khandelwal (2011) for India. See also the discussion in Holmes and Schmitz (2010).

Agreement and similarly exploits the heterogeneity in affected sectors. He finds productivity increases in excess of 15 percent for both shrinking (that is, import competing) sectors as well as expanding (exporting) sectors.

Khandelwal, Schott, and Wei (2013) study another specific episode of trade reform—the elimination of export quotas on Chinese textile and clothing by the United States, the European Union, and Canada in 2005. While export quotas allocated via market arrangements generate standard misallocation effects on aggregate productivity, their empirical analysis shows that the quota removal generated larger effects because the government had allocated quotas to less-productive state-owned enterprises. They find that more than 70 percent of the overall productivity gain is due to quota misallocation whereas the remaining 30 percent is due to standard misallocation from eliminating the quotas.

Trade policy may also affect misallocation via its effect on competition, which is often proxied by markups. Edmond, Midrigan, and Xu (2015) calibrate a model to Taiwanese manufacturing data and find that moving from autarky to free trade decreases markup heterogeneity and leads to an increase in total factor productivity of slightly more than 12 percent.[10]

Financial and Informational Frictions

Financial market imperfections are perhaps the single most studied source of misallocation. The positive correlation between financial market development and output per capita is a robust empirical finding (Levine 1997). The literature on financial market development and economic development is too large to discuss in any detail (for a survey of the broader related literature on financial development, see Buera, Kaboski, and Shin 2015). We focus on papers in this literature that have quantified the misallocation of capital across producers due to credit constraints. This literature has generated a range of estimates, some of them quite large.

Consistent with our earlier warning about the importance of model features, it is now well understood that the effects depend in an important way on such features, specifically the scope for individuals to accumulate assets in order to grow out of financial constraints. This in turn is heavily influenced by the persistence of productivity (or demand) at the producer level. As the literature has made more attempts to model this feature and discipline it using microdata, the resulting effects of capital misallocation on total factor productivity have diminished. For example, Midrigan and Xu (2014) find that the magnitude of this effect is no more than about 10 percent (see also Buera, Kaboski, and Shin 2011; Greenwood, Sanchez, and Wang 2013; Moll 2014). Gopinath et al. (2015) found that a large part of the increased misallocation of capital in Mediterranean countries after 1999 is accounted for by financial frictions, but the magnitude of the effect is on the order of a 3 percent drop in total factor productivity.

[10] Epifani and Gancia (2011) show that dispersion of markups across manufacturing industries is significantly greater in poorer countries than in richer countries, but did not assess what this implies for cross-country differences in productivity.

Other relevant market frictions include imperfect information, imperfect insurance, and imperfect enforcement of contracts. For example, David, Hopenhayn, and Venkateswaran (2016) identify information frictions by combining production and stock market data of firms and find that these types of frictions can reduce aggregate productivity by 7–10 percent in China and India. Imperfect insurance and credit restrictions have also played a prominent role in development economics (Udry 2012).[11] Caselli and Gennaioli (2013) study the effects of poor contract enforcement as it affects management of family-run firms, and show that the effects on aggregate total factor productivity can be substantial.

Summary

Studies using the direct approach often find sources of misallocation that reduce total factor productivity, but even taken together, the effects from these studies are small compared to the indirect effects noted earlier. One possibility is that the indirect effects estimated earlier are overestimates of the extent of differences in misallocation. Alternatively, it is possible that the aggregate effects are the result of many different individual factors, each of which contributes a small part, so that we will never isolate a single dominant factor. Or perhaps the existing analyses of direct effects, based on relatively simple models and somewhat generic treatments of potential sources of misallocation, may not adequately capture the full extent of frictions that are present in less-developed counties.

Additional Consequences of Misallocation

The policies and institutions that distort firm-level choices of labor and capital at a point in time, thereby generating misallocation, are also likely to affect entry and exit decisions as well as firm-level investments that influence future productivity. These effects operate via the selection and technology channels discussed earlier and represent consequences beyond those estimated using the indirect method.

A growing body of work emphasizes the broader consequences of misallocation in settings with selection and/or technology effects. All of the previously noted empirical studies of trade liberalizations using producer-level data find an important role for both selection effects and producer-level productivity gains. Bustos (2011) specifically finds that producers in Argentina invest more in technology upgrading in response to trade liberalizations.[12] Selection effects are featured prominently in the theoretical analysis of Melitz (2003). More recently

[11] Munshi and Rosenzweig (2016) emphasize risk and differential insurance arrangements between rural and urban sectors in restricting labor mobility, therefore potentially generating labor misallocation across space.

[12] Bloom, Draca, and Van Reenen (2016) provide similar evidence for firms in Europe. Aw, Roberts, and Xu (2011) estimate a structural model of trade and research and development investment using data on Taiwanese electronics producers. In simulations, they find that trade liberalizations increase producer-level productivity via increased investment in research and development.

these models have been extended to allow for endogenous plant-level productivity responses as well (for examples, see Costantini and Melitz 2008; Caliendo and Rossi-Hansberg 2012; Rubini 2014; Mayer, Melitz, and Ottaviano 2017). In the financial frictions literature, the bulk of productivity effects are due to distorted occupational choice decisions (highly productive entrepreneurs that do not operate, as in Buera, Kaboski, and Shin 2011) and technology investment (Midrigan and Xu 2014). In the agricultural sector, land institutions that prevent the reallocation of land to best uses also act as a deterrent for highly productive farmers who may instead choose to work outside of agriculture (Adamopoulos et al. 2017). In the context of labor market regulations, Da-Rocha, Tavares, and Restuccia (2016) study the effect of firing costs on productivity in a model that includes an endogenous choice for innovation, and find that the dynamic effects on productivity are substantial, increasing the total factor productivity loss from around 2 percent due to static misallocation to an overall effect of 4 percent, for firing costs equivalent to one year's wages. Peters (2016) studies a model of innovation in which limited competition leads to heterogeneity in markups, and shows that the dynamic effect of markup heterogeneity is more than four times larger than the static misallocation effects.

From a modeling point of view, the key issue is to extend the simple static model of heterogeneous producers that we outlined earlier to a dynamic setting that includes endogenous decisions that influence future productivity. Restuccia (2013) provides an early example of using such a model to analyze the consequences of hypothetical distortions. He assumes there are upfront investments in productivity when a new establishment is created, and higher investments yield higher-productivity establishments in expectation. In this setting, implicit taxes on higher-productivity establishments lower the incentive for investments that are expected to raise productivity and hence lower the overall distribution of establishment-level productivities. He uses this framework to shed light on the productivity gap between Latin America and the United States.[13] Another recent paper along these lines is Hsieh and Klenow (2014) on the life cycle of manufacturing plants in India, Mexico, and the United States. Their analysis is motivated by the empirical observation that older plants in India and Mexico are much less productive relative to young plants than is the case in the United States. Given this difference in relative productivities, it is efficient that older plants in India and Mexico are relatively small compared to their counterparts in the United States. They show that, in analyses including life-cycle investment in productivity improvements at the establishment level, the greater implicit taxes faced by more-productive establishments in India and Mexico can potentially account for a large share of the differences in productivity gradients with age across plants.

Bento and Restuccia (forthcoming) build a model that allows for productivity investments both at the time of entry as well as along the life cycle post-entry.

[13] Many other contributions have recognized the feedback from misallocation to the determination of firm-level productivity levels; see Hopenhayn (2016), Da-Rocha, Tavares, and Restuccia (2017), and the references therein.

They find that the greater implicit taxes faced by more-productive establishments in India compared to the United States reduces aggregate productivity in India by 53 percent and average establishment size by 86 percent. They decompose this productivity effect into three components: a static effect of misallocation, a life-cycle effect due to lower life-cycle investment in productivity, and an entry productivity effect capturing the effect of lower investment in productivity at the time of entry. The reduction in aggregate productivity is roughly equally shared between static misallocation and entry-level productivity investments. In their model, life-cycle investment in productivity plays a minor role because the reduction in life-cycle productivity growth is offset by its effect on establishment entry.

In related work, Ayerst (2016) attempts to connect misallocation with barriers to technology adoption and diffusion lags across countries, based on the insight that policies and institutions that generate misallocation may create disincentives to adopt the most modern and best technologies. Bigio and La'O (2016) study the effect of policy distortions in an environment with production networks as emphasized in the survey article of Jones (2013). They find that the productivity effects of policy distortions in a model with production networks are roughly four times that in the model of the economy that abstracts from the network structure.

Overall, the work just described suggests that studies of misallocation should look for opportunities to go beyond static effects of misallocation, and focus on the potentially much larger dynamic effects. We believe that micro-level panel data will be critical to producing compelling empirical evidence about these channels.

Where to from Here?

To take stock, we revisit the three questions posed in the introduction.

First, how important is misallocation? Misallocation appears to be a substantial channel in accounting for productivity differences across countries, but the measured magnitude of the effects depends on the approach and context. Productivity losses from misallocation reported using the indirect approach are typically an order of magnitude or more larger than the losses associated with specific policies and institutions reported using the direct approach. More work is needed on the various mechanisms that can potentially amplify the effect of misallocation on aggregate productivity and in particular in connecting policies that generate misallocation with observed micro productivity distributions.

Second, what are the causes of misallocation? The research has not found a dominant source of misallocation; instead, many specific factors seem to contribute a small part of the overall effect. Our view is that studies that follow the direct approach are more likely to reach concrete, persuasive, and specific conclusions of practical policy relevance. However, the indirect approach can be especially valuable in diagnosing important dimensions of misallocation: for example, whether it is more significant in some sectors, or whether it is related to specific factors of production such as capital, labor, or land.

Third, are there additional costs to misallocation? The answer is clearly "yes," and whereas much of the literature has focused on static misallocation, we think the dynamic effects of misallocation deserve much more attention going forward.

In moving ahead, we expect that the increasing availability of micro datasets, especially firm-level panel datasets, is likely to yield opportunities to exploit changes in policies and institutions and variations across individuals, firms, regions, and other relevant dimensions, and will offer new opportunities to study the role of misallocation.

We are also intrigued by aspects of misallocation that reach beyond the issues of how labor and capital might be misallocated across firms. For example, discrimination, culture, and social norms can lead to misallocation of talent across employment status, occupations, and sectors. Hnatkovska, Lahiri, and Paul (2012) document the misallocation of talent in India that arises as a result of the caste system, and document that these barriers have decreased dramatically over the last 20 years. In a similar spirit, Hsieh, Hurst, Jones, and Klenow (2015) discuss shifts in the allocation of talent across occupations in the United States. For example, in 1960 around 94 percent of doctors and lawyers were white men, whereas by 2008, the share declined to 62 percent. Given that innate talent is unlikely to feature such a concentration across gender and races, the occupational distribution in 1960 reflects misallocation of talent and the observed convergence represents an improvement in the allocation. They estimate that convergence in the occupational distribution across races and gender can account for 15 to 20 percent of growth in aggregate output per worker in the United States between 1960 and 2008. We think this work suggests a promising direction for additional research on the allocation of talent and how it differs across economies.

■ *The authors thank the editors Enrico Moretti, Gordon Hanson, and Timothy Taylor for useful comments.*

References

Adamopoulos, Tasso, Loren Brandt, Jessica Leight, and Diego Restuccia. 2017. "Misallocation, Selection and Productivity: A Quantitative Analysis with Panel Data from China." NBER Working Paper 23039.

Adamopoulos, Tasso, and Diego Restuccia. 2014. "The Size Distribution of Farms and International Productivity Differences." *American Economic Review* 104(6): 1667–97.

Adamopoulos, Tasso, and Diego Restuccia. 2015. "Land Reform and Productivity: A Quantitative Analysis with Micro Data." University of Toronto Department of Economics Working Paper 540.

Asker, John, Allan Collard-Wexler, and Jan De Loecker. 2014. "Dynamic Inputs and Resource (Mis)allocation." *Journal of Political Economy* 122(5): 1013–63.

Aw, Bee Yan, Mark J. Roberts, and Daniel Yi Xu. 2011. "R&D Investment, Exporting, and Productivity Dynamics." *American Economic Review* 101(4): 1312–44.

Ayerst, Stephen. 2016. "Idiosyncratic Distortions and Technology Adoption." University of Toronto Department of Economics Working Paper 571.

Baily, Martin Neil, Charles Hulten, and David Campbell. 1992. "Productivity Dynamics in Manufacturing Plants." *Brookings Papers on Economic Activity*: 187–267.

Banerjee, Abhijit. 1999. "Land Reforms: Prospects and Strategies." Massachusetts Institute of Technology (MIT) Department of Economics Working Paper 99–24.

Banerjee, Abhijit V., and Esther Duflo. 2005. "Growth Theory through the Lens of Development Economics." In *Handbook of Economic Growth,* Vol. 1, Part A, edited by Philippe Aghion and Steven N. Durlauf, 473–552. North Holland.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta. 2013. "Cross-Country Differences in Productivity: The Role of Allocation and Selection." *American Economic Review* 103(1): 305–34.

Bento, Pedro, and Diego Restuccia. Forthcoming. "Misallocation, Establishment Size, and Productivity." *American Economic Journal: Macroeconomics.*

Bernard, Andrew B., J. Bradford Jensen, and Peter K. Schott. 2006. "Trade Costs, Firms and Productivity." *Journal of Monetary Economics* 53(5): 917–37.

Besley, Timothy, and Maitreesh Ghatak. 2010. "Property Rights and Economic Development." In *Handbook of Development Economics. Volume 5*, edited by Dani Rodrik and Mark Rosenzweig, 4525–95. North Holland.

Bigio, Saki, and Jennifer La'O. 2016. "Financial Frictions in Production Networks." NBER Working Paper 22212.

Bils, Mark, Peter J. Klenow, and Cian Ruane. 2017. "Misallocation or Mismeasurement?" http://www.klenow.com/misallocation-mismeasurement-paper.pdf.

Bloom, Nicholas. 2009. "The Impact of Uncertainty Shocks." *Econometrica* 77(3): 623–85.

Bloom, Nicholas, Mirko Draca, and John Van Reenen. 2016. "Trade Induced Technical Change? The Impact of Chinese Imports on Innovation, IT and Productivity." *Review of Economic Studies* 83(1): 87–117.

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

Bollard, Albert, Peter J. Klenow, and Gunjan Sharma. 2013. "India's Mysterious Manufacturing Miracle." *Review of Economic Dynamics* 16(1): 59–85.

Bond, Stephen, and John Van Reenen. 2007. "Microeconometric Models of Investment and Employment." In *Handbook of Econometrics,* Vol. 6, Part A, edited by James J. Heckman and Edward E. Leamer, 4417–98. North Holland.

Brandt, Loren, Trevor Tombe, and Xiaodong Zhu. 2013. "Factor Market Distortions across Time, Space and Sectors in China." *Review of Economic Dynamics* 16(1): 39–58.

Buera, Francisco J., Joseph P. Kaboski, and Yongseok Shin. 2011. "Finance and Development: A Tale of Two Sectors." *American Economic Review* 101(5): 1964–2002.

Buera, Francisco J., Joseph P. Kaboski, and Yongseok Shin. 2015. "Entrepreneurship and Financial Frictions: A Macro-development Perspective." *Annual Review of Economics* 7(1): 409–36.

Busso, Matías, María Victoria Fazio, and Santiago Levy. 2012. "(In)formal and (Un)productive: The Productivity Costs of Excessive Informality in Mexico." Inter-American Development Bank Working Paper IDB-WP-341.

Busso, Matías, Lucía Madrigal, and Carmen Pagés. 2013. "Productivity and Resource Misallocation in Latin America." *B.E. Journal of Macroeconomics* 13(1): 903–32.

Bustos, Paula. 2011. "Trade Liberalization, Exports, and Technology Upgrading: Evidence on the Impact of MERCOSUR on Argentinian Firms." *American Economic Review* 101(1): 304–40.

Caballero, Ricardo J., Takeo Hoshi, and Anil K. Kashyap. 2008. "Zombie Lending and Depressed Restructuring in Japan." *American Economic Review* 98(5): 1943–77.

Caliendo, Lorenzo, and Fernando Parro. 2015. "Estimates of the Trade and Welfare Effects of NAFTA." *Review of Economic Studies* 82(1): 1–44.

Caliendo, Lorenzo, and Esteban Rossi-Hansberg. 2012. "The Impact of Trade on Organization and Productivity." *Quarterly Journal of Economics* 127(3): 1393–1467.

Calligaris, Sara. 2015. "Misallocation and Total Factor Productivity in Italy: Evidence from Firm-Level Data." *Labour* 29(4): 367–93.

Caselli, Francesco. 2005. "Accounting for Cross-Country Income Differences." *Handbook of Economic Growth,* Vol. 1, Part A, edited by Philippe Aghion and Steven N. Durlauf, 679–741. North Holland.

Caselli, Francesco, and Nicola Gennaioli. 2013. "Dynastic Management." *Economic Inquiry* 51(1): 971–96.

Chen, Chaoran, Diego Restuccia, and Raül Santaeulàlia-Llopis. 2017. "Land Markets, Resource Allocation, and Agricultural Productivity." https://www.economics.utoronto.ca/diegor/research/CRS_paper.pdf.

Chen, Kaiji, and Alfonso Irarrazabal. 2015. "The Role of Allocative Efficiency in a Decade of Recovery." *Review of Economic Dynamics* 18(3): 523–50.

Cooper, Russell W., and John C. Haltiwanger. 2006. "On the Nature of Capital Adjustment Costs." *Review of Economic Studies* 73(3): 611–33.

Costantini, James A., and Marc J. Melitz. 2008. "The Dynamics of Firm-Level Adjustment to Trade Liberalization." In *The Organization of Firms in a Global Economy*, edited by Elhanan Helpman, Dalia Marin, and Thierry Verdier, 107–41. Harvard University Press.

Da-Rocha, José-María, Marina Mendes Tavares, and Diego Restuccia. 2016. "Firing Costs, Misallocation, and Aggregate Productivity." NBER Working Paper 23008.

Da-Rocha, José-María, Marina Mendes Tavares, and Diego Restuccia. 2017. "Policy Distortions and Aggregate Productivity with Endogenous Establishment-Level Productivity." University of Toronto Department of Economics Working Paper 579.

David, Joel M., Hugo A. Hopenhayn, and Venky Venkateswaran. 2016. "Information, Misallocation, and Aggregate Productivity." *Quarterly Journal of Economics* 131(2): 943–1005.

David, Joel M., and Venky Venkateswaran. 2017. "Capital Misallocation: Frictions or Distortions?" NBER Working Paper 23129.

de Janvry, Alain. 1981. "The Role of Land Reform in Economic Development: Policies and Politics." *American Journal of Agricultural Economics* 63(2): 384–92.

de Janvry, Alain, Kyle Emerick, Marco Gonzalez-Navarro, and Elisabeth Sadoulet. 2015. "Delinking Land Rights from Land Use: Certification and Migration in Mexico." *American Economic Review* 105(10): 3125–49.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2008. "Returns to Capital in Microenterprises: Evidence from a Field Experiment." *Quarterly Journal of Economics* 123(4): 1329–72.

de Vries, Gaaitzen J. 2014. "Productivity in a Distorted Market: The Case of Brazil's Retail Sector." *Review of Income and Wealth* 60(3): 499–524.

D'Erasmo, Pablo N., and Hernan J. Moscoso Boedo. 2012. "Financial Structure, Informality and Development." *Journal of Monetary Economics* 59(3): 286–302.

Deininger, Klaus, and Gershon Feder. 2001. "Land Institutions and Land Markets." In *Handbook of Agricultural Economics,* vol. 1A: *Agricultural Production*, edited by Bruce L. Gardner and Gordon C. Rausser, 287–331. North Holland.

Dias, Daniel, Carlos Robalo Marques, and Christine Richmond. 2016a. "Comparing Misallocation between Sectors in Portugal." Banco de Portugal, Economics Research Department Economic Bulletin and Financial Stability Report Article 201602.

Dias, Daniel A., Carlos Robalo Marques, and Christine Richmond. 2016b. "Misallocation and Productivity in the Lead up to the Eurozone Crisis." *Journal of Macroeconomics* 49: 46–70.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70(5): 1741–79.

Eaton, Jonathan, Samuel Kortum, and Francis Kramarz. 2011. "An Anatomy of International Trade: Evidence from French Firms." *Econometrica* 79(5): 1453–98.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu. 2015. "Competition, Markups, and the Gains from International Trade." *American Economic Review* 105(10): 3183–221.

Epifani, Paolo, and Gino Gancia. 2011. "Trade, Markup Heterogeneity and Misallocations." *Journal of International Economics* 83(1): 1–13.

Fajgelbaum, Pablo D., Eduardo Morales, Juan Carlos Suárez Serrato, and Owen M. Zidar. 2015. "State Taxes and Spatial Misallocation." NBER Working Paper 21760.

Fernandes, Ana M. 2007. "Trade Policy, Trade Volumes, and Plant-Level Productivity in Colombian Manufacturing Industries." *Journal of International Economics* 71(1): 52–71.

Foster, Lucia, John Haltiwanger, and Chad Syverson. 2008. "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review* 98(1): 394–425.

Fujii, Daisuke, and Yoshio Nozawa. 2013. "Misallocation of Capital During Japan's Lost Two Decades." Development Bank of Japan Working Paper 1304.

Garcia-Santana, Manuel, and Josep Pijoan-Mas. 2014. "The Reservation Laws in India and the Misallocation of Production Factors." *Journal of Monetary Economics* 66: 193–209.

Garicano, Luis, Claire Lelarge, and John Van Reenen. 2016. "Firm Size Distortions and the Productivity Distribution: Evidence from France." *American Economic Review* 106(11): 3439–79.

Gollin, Douglas, Stephen Parente, and Richard Rogerson. 2002. "The Role of Agriculture in Development." *American Economic Review* 92(2): 160–64.

Gopinath, Gita, Sebnem Kalemli-Ozcan, Loukas Karabarbounis, and Carolina Villegas-Sanchez. 2015. "Capital Allocation and Productivity in South Europe." NBER Working Paper 21453.

Gourio, François, and Nicolas Roys. 2014. "Size-Dependent Regulations, Firm Size Distribution, and Reallocation." *Quantitative Economics* 5(2): 377–416.

Greenwood, Jeremy, Juan M. Sanchez, and Cheng Wang. 2013. "Quantifying the Impact of Financial Development on Economic Development." *Review of Economic Dynamics* 16(1): 194–215.

Guner, Nezih, Gustavo Ventura, and Yi Xu. 2008. "Macroeconomic Implications of Size-Dependent Policies." *Review of Economic Dynamics* 11(4): 721–44.

Hall, Robert E., and Charles I. Jones. 1999. "Why Do Some Countries Produce So Much More Output per Worker than Others?" *Quarterly Journal of Economics* 114(1): 83–116.

Heckman, James J., and Carmen Pagés. 2004. "Introduction." In *Law and Employment: Lessons from Latin America and the Caribbean*, edited by James J. Heckman and Carmen Pagés, 1–108. University of Chicago Press.

Hnatkovska, Viktoria, Amartya Lahiri, and Sourabh Paul. 2012. "Castes and Labor Mobility." *American Economic Journal: Applied Economics* 4(2): 274–307.

Holmes, Thomas J., and James A. Schmitz Jr. 2010. "Competition and Productivity: A Review of Evidence." *Annual Review of Economics* 2(1): 619–42.

Hopenhayn, Hugo A. 1992. "Entry, Exit, and Firm Dynamics in Long Run Equilibrium." *Econometrica* 60(5): 1127–50.

Hopenhayn, Hugo A. 2014. "Firms, Misallocation, and Aggregate Productivity: A Review." In *Annual Review of Economics*, Vol. 6, edited by Kenneth J. Arrow and Timothy F. Bresnahan, 735–70. Annual Reviews.

Hopenhayn, Hugo A. 2016. "Firm Size and Development." *Economía* 17(1): 27–49.

Hopenhayn, Hugo A., and Richard Rogerson. 1993. "Job Turnover and Policy Evaluation: A General Equilibrium Analysis." *Journal of Political Economy* 101(5): 915–38.

Hsieh, Chang-Tai, Erik Hurst, Charles I. Jones, and Peter J. Klenow. 2013. "The Allocation of Talent and U.S. Economic Growth." NBER Working Paper 18693.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *Quarterly Journal of Economics* 124(4): 1403–48.

Hsieh, Chang-Tai, and Peter J. Klenow. 2014. "The Life Cycle of Plants in India and Mexico." *Quarterly Journal of Economics* 129(3): 1035–84.

Hsieh, Chang-Tai, and Enrico Moretti. 2015. "Why Do Cities Matter? Local Growth and Aggregate Growth." NBER Working Paper 21154.

Jones, Charles I. 2013. "Misallocation, Input–Output Economics, and Economic Growth." In *Advances in Economics and Econometrics: Tenth World Congress. Volume II: Applied Economics*, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 419–56. Cambridge University Press.

Jones, Charles I., 2016. "The Facts of Economic Growth." In *Handbook of Macroeconomics. Volume 2*, edited by John B. Taylor and Harald Uhlig, 3–69. North Holland.

Kalemli-Ozcan, Sebnem, and Bent E. Sørensen. 2016. "Misallocation, Property Rights, and Access to Finance: Evidence from Within and Across Africa." Chap. 5 in *African Successes: Modernization and Development*, Vol. 3, edited by Sebastian Edwards, Simon Johnson, and David N. Weil. University of Chicago Press.

Kehoe, Timothy, Pau S. Pujolàs, and Jack Rossbach. Forthcoming. "Quantitative Trade Models: Developments and Challenges." *Annual Reviews of Economics*.

Khandelwal, Amit K., Peter K. Schott, and Shang-Jin Wei. 2013. "Trade Liberalization and Embedded Institutional Reform: Evidence from Chinese Exporters." *American Economic Review* 103(6): 2169–95.

Klenow, Peter J., and Andres Rodriguez-Clare. 1997. "The Neoclassical Revival in Growth Economics: Has It Gone Too Far?" In *NBER Macroeconomics Annual 1997*, edited by Ben S. Bernanke and Julio J. Rotemberg, 73–103. MIT Press.

La Porta, Rafael, and Andrei Shleifer. 2014. "Informality and Development." *Journal of Economic Perspectives* 28(3): 109–26.

Lagos, Ricardo. 2006. "A Model of TFP." *Review of Economic Studies* 73(4): 983–1007.

Leal Ordóñez, Julio Cesar. 2014. "Tax Collection, the Informal Sector, and Productivity." *Review of Economic Dynamics* 17(2): 262–86.

Levine, Ross. 1997. "Financial Development and Economic Growth: Views and Agenda." *Journal of Economic Literature* 35(2): 688–726.

Lewis, W. A. 1954. "Economic Development with Unlimited Supplies of Labour." *Manchester School of Economic and Social Studies* 22: 139–91.

Mayer, Thierry, Marc Melitz, and Gianmacro I. P. Ottaviano. 2016. "Product Mix and Firm Productivity Responses to Trade Competition." NBER Working Paper 22433.

Melitz, Marc J. 2003. "The Impact of Trade on Intra-industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71(6): 1695–725.

Melitz, Marc J., and Stephen Redding. 2014. "Heterogeneous Firms and Trade." In *Handbook of International Economics*, Vol. 4, edited by Gita Gopinath, Elhanan Helpman, and Ken Rogoff, 1–54. New Holland.

Midrigan, Virgiliu, and Daniel Yi Xu. 2014. "Finance and Misallocation: Evidence from Plant-Level Data." *American Economic Review* 104(2): 422–58.

Moll, Benjamin. 2014. "Productivity Losses from Financial Frictions: Can Self-Financing Undo Capital Misallocation?" *American Economic Review* 104(10): 3186–221.

Munshi, Kaivan, and Mark Rosenzweig. 2016. "Networks and Misallocation: Insurance, Migration, and the Rural-Urban Wage Gap." *American Economic Review* 106(1): 46–98.

Oberfield, Ezra. 2013. "Productivity and Misallocation during a Crisis: Evidence from the Chilean Crisis of 1982." *Review of Economic Dynamics* 16(1): 100–119.

Olley, G. Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica* 64(6): 1263–97.

Pavcnik, Nina. 2002. "Trade Liberalization, Exit, and Productivity Improvement: Evidence from Chilean Plants." *Review of Economic Studies* 69(1): 245–76.

Peters, Michael. 2016. "Heterogeneous Markups, Growth and Endogenous Misallocation." Unpublished paper.

Prescott, Edward C. 1998. "Needed: A Theory of Total Factor Productivity." *International Economic Review* 39(3): 525–51.

Rauch, James E. 1991. "Modelling the Informal Sector Formally." *Journal of Development Economics* 35(1): 33–47.

Reis, Ricardo. 2013. "The Portuguese Slump and Crash and the Euro Crisis." *Brookings Papers on Economic Activity* 46(1): 143–93.

Restuccia, Diego. 2013. "The Latin American Development Problem: An Interpretation." *Economia: Journal of the Latin American and Caribbean Economic Association* 13(2): 69–100.

Restuccia, Diego, Dennis Tao Yang, and Xiaodong Zhu. 2008. "Agriculture and Aggregate Productivity: A Quantitative Cross-Country Analysis." *Journal of Monetary Economics* 55(2): 234–50.

Restuccia, Diego, and Richard Rogerson. 2008. "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments." *Review of Economic Dynamics* 11(4): 707–20.

Restuccia, Diego, and Richard Rogerson. 2013. "Misallocation and Productivity: Editorial." *Review of Economic Dynamics* 16(1): 1–10.

Restuccia, Diego, and Raul Santaeulalia-Llopis. 2017. "Land Misallocation and Productivity." NBER Working Paper 23128.

Rubini, Loris. 2014. "Innovation and the Trade Elasticity." *Journal of Monetary Economics* 66: 32–46.

Sandleris, Guido, and Mark L. J. Wright. 2014. "The Costs of Financial Crises: Resource Misallocation, Productivity, and Welfare in the 2001 Argentine Crisis." *Scandinavian Journal of Economics* 116(1): 87–127.

Song, Zheng, Kjetil Storesletten, and Fabrizio Zilibotti. 2011. "Growing Like China." *American Economic Review* 101(1): 196–233.

Syverson, Chad. 2011. "What Determines Productivity?" *Journal of Economic Literature* 49(2): 326–65.

Tombe, Trevor. 2015. "The Missing Food Problem: Trade, Agriculture, and International Productivity Differences." *American Economic Journal: Macroeconomics* 7(3): 226–58.

Tombe, Trevor, and Xiaodong Zhu. 2015. "Trade, Migration and Productivity: A Quantitative Analysis of China." University of Toronto Department of Economics Working Paper 542.

Topalova, Petia, and Amit Khandelwal. 2011. "Trade Liberalization and Firm Productivity: The Case of India." *Review of Economics and Statistics* 93(3): 995–1009.

Trefler, Daniel. 2004. "The Long and Short of the Canada–U.S. Free Trade Agreement." *American Economic Review* 94(4): 870–95.

Udry, Christopher. 2012. "Misallocation, Growth and Financial Market Imperfections: Microeconomic Evidence." Lecture presented at the annual meeting of the Society for Economic Dynamics, Limassol. https://www.economicdynamics.org/wp-content/uploads/SED_Annual_Meeting/udry2012.pdf.

Waugh, Michael E. 2010. "International Trade and Income Differences." *American Economic Review* 100(5): 2093–124.

Ziebarth, Nicolas L. 2013. "Are China and India Backward? Evidence from the 19th Century U.S. Census of Manufactures." *Review of Economic Dynamics* 16(1): 86–99.

Ziebarth, Nicolas L. 2015. "The Great Depression through the Eyes of the Census of Manufactures." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 48(4): 185–94.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 175–194

Federal Budget Policy with an Aging Population and Persistently Low Interest Rates

Douglas W. Elmendorf and Louise M. Sheiner

The federal budget is on an unsustainable path. Federal debt has surged in the past decade and is now larger relative to gross domestic product (GDP) than at any time in US history except for the period around the end of World War II. Moreover, the Congressional Budget Office (2016b) projects that debt will rise substantially further relative to GDP if current laws and policies are not changed, increasing from about 75 percent today to about 140 percent 30 years from now, as shown in Figure 1. For comparison, federal debt averaged 39 percent of GDP during the past 50 years.

Some observers have argued that the projections for high and rising debt pose a grave threat to the country's economic future and also mean that the government has less fiscal space to respond to recessions or other unexpected developments, so they urge significant changes in tax or spending policies to reduce federal borrowing. In stark contrast, other observers have noted that interest rates on long-term federal debt are extremely low and have argued that such persistently low interest rates justify additional federal borrowing and investment, at least for the short and medium term.

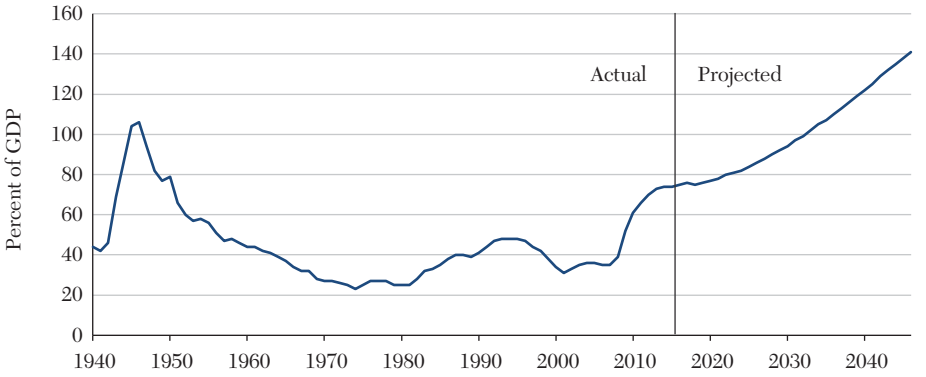
Our analysis of this controversy focuses on two main issues: the aging of the US population and interest rates on US government debt. It is generally understood that these factors play an important role in the projected path of the US

■ *Douglas W. Elmendorf is Dean and Don K. Price Professor of Public Policy, Harvard Kennedy School, Cambridge, Massachusetts. From January 2009 through March 2015, he served as Director of the Congressional Budget Office. Louise M. Sheiner is Senior Fellow in Economic Studies and Policy Director, Hutchins Center on Fiscal and Monetary Policy, both at the Brookings Institution, Washington, DC. Their email addresses are Doug_Elmendorf@hks.harvard.edu and lsheiner@brookings.edu.*

†For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.175

doi=10.1257/jep.31.3.175

Figure 1
Federal Debt Held by the Public



Source: Congressional Budget Office (2016b).

debt-to-GDP ratio. What is less recognized is that these changes also have implications for the *appropriate* level of US debt.

Economists and policymakers have anticipated for some time that rapid growth in the share of Americans over age 65 would sharply raise spending for Social Security, Medicare, and certain other federal programs relative to GDP. Indeed, population aging and a projected increase in per-capita healthcare spending now explain more than all of the projected growth in noninterest federal spending over the next few decades. However, population aging also reduces the share of the population in the labor force, which lowers feasible consumption relative to what it would be otherwise; we show in this paper that the optimal social response to population aging would be higher national saving—the sum of private saving and public saving (or dissaving)—over the next decade equal to about 1 percent of GDP.

Both market readings and detailed analyses by a number of researchers suggest that Treasury interest rates are likely to remain well below their historical norms for years to come, which represents a sea change for budget policy. We argue that many—though not all—of the factors that may be contributing to the historically low level of interest rates imply that both federal debt and federal investment should be substantially larger than they would be otherwise.

We conclude that, although significant policy changes to reduce federal budget deficits ultimately will be needed, they do not have to be implemented right away. Instead, the focus of federal budget policy over the coming decade should be to increase federal investment while enacting changes in federal spending and taxes that will reduce deficits gradually over time.

Implications of the Projected Increase in Federal Debt

Stabilizing federal debt relative to GDP would require substantial changes in policies; returning the debt-to-GDP ratio to its historical average would require

even larger changes. Holding down debt relative to GDP would have significant economic advantages in the long run.

The Projected Path of Federal Debt

The so-called baseline projections of the Congressional Budget Office are conditioned on the assumption that current law (generally) persists and that annual appropriations (which currently account for one-third of noninterest federal spending) grow at the same pace as GDP after the next 10 years. (For a full description of the policy assumptions underlying the baseline projections, see CBO 2016b.)

In those projections, federal deficits rise to nearly 5 percent of GDP by 2026, and federal debt held by the public reaches 86 percent of GDP that year.[1] By 2046, debt is projected to reach 141 percent of GDP and to be on a rising trajectory. The driving factor behind the projected run-up in deficits and debt is growth in federal spending for older Americans and for health care that is not fully offset by reductions in other spending or increases in revenues. In particular, noninterest spending is projected to increase from nearly 20 percent of GDP today to more than 22 percent by 2046. By our calculations, the aging of the population alone will more than account for that rise, generating an increase in Social Security and Medicare spending relative to GDP of roughly 2½ percentage points during that period (Elmendorf and Sheiner 2016). In addition, growth in federal healthcare spending per beneficiary will almost certainly exceed growth in GDP per person, as healthcare spending throughout the US economy has done for many decades. Meanwhile, projected revenues edge up from a little more than 18 percent of GDP in 2016 to a little more than 19 percent in 2046, primarily because growth in inflation-adjusted income will push taxpayers into higher tax brackets. Growing noninterest deficits are accompanied by growing interest payments, which are projected to jump from about 1.5 percent of GDP to nearly 6 percent, owing to increases in both debt and interest rates.

Of course, such projections are highly uncertain. Future productivity, interest rates, healthcare costs, life expectancy, and other factors might differ from current forecasts. As one illustration of this uncertainty, the Congressional Budget Office's projections of debt under alternative assumptions about the key demographic and economic inputs to its projections range between 93 percent and 196 percent of GDP in 2046. In addition, the baseline projections incorporate some changes in policies that may prove politically unpalatable. For example, nondefense appropriations have shown no trend relative to GDP during the past 50 years—perhaps because many items they cover, like highways, have demands that grow with GDP. However, those appropriations are currently constrained by statutory caps first enacted in 2011, and those caps imply that such appropriations will be smaller relative to GDP in each year after 2018 than in any year in that earlier half-century. If nondefense appropriations are ultimately increased relative to that projection, then larger changes in other tax and spending policies will be needed to put the budget on a sustainable path.

[1] Congressional Budget Office (2010) addresses broader measures of the federal government's financial position. Those measures tend to follow roughly the same contour as debt held by the public.

Many other nations also face the challenge of high and rising public debt, and the budgetary pressure of population aging is an important factor for many of them. According to data from the International Monetary Fund (2016) on government debt net of financial assets for all levels of government in 2015, US debt equaled 80 percent of GDP, and several other countries were in very similar positions: France, at 88 percent; Spain, 80 percent; and the United Kingdom, 80 percent. Those figures are high relative to the historical experience of all of those countries. By this measure, current debt burdens are smaller relative to the United States in Germany (48 percent) and Canada (26 percent) but larger in Japan (125 percent) and Italy (113 percent).[2]

Although no one knows how much debt is "too much," debt cannot increase indefinitely relative to GDP, as it almost surely would in the United States (and many other countries) without significant changes in spending and tax policies. Therefore, policy changes will be needed at some point.

Optimal Policy in Response to Rising Debt

According to the Congressional Budget Offices's (2016b) projections, the debt-to-GDP ratio in 30 years would equal the ratio today if the country adopted immediate and permanent spending cuts, tax increases, or a combination of both equal to 1¾ percent of GDP. With current US GDP equal to roughly $18 trillion, such changes would amount to $315 billion today and would grow with GDP. Returning the debt-to-GDP ratio in 30 years to its earlier 50-year average could be achieved through immediate, permanent policy changes equal to almost 3 percent of GDP, or a combination of $540 billion in spending cuts and tax increases today. Of course, if the changes were phased in slowly, even larger changes would be needed later.

The most common argument for holding down the debt-to-GDP ratio is that doing so would lead to greater national savings in the long run, and the higher level of savings would lead to more capital, higher productivity, and higher wages and incomes. In the Congressional Budget Office's modeling, each $1 reduction in federal debt raises national saving by 57 cents in the long run—an amount that is less than a dollar because private saving would be diminished by the decline in interest rates that would result from less debt and by other factors (CBO 2014a). However, the tax increases or spending cuts that would reduce federal debt would lower output and private saving in the short run if the Federal Reserve was unable to reduce interest rates enough to offset the contractionary change in fiscal policy. If hysteresis effects on output and employment are significant—that is, if short-term changes in aggregate demand generate lasting changes in the supply of labor, capital, or technological progress, and therefore sustainable output and employment—then the "short run" might have echoes over time.

Another argument for holding down the debt-to-GDP ratio is that doing so would put the government in a better position to deal with unexpected developments. For example, if an economic downturn warranted an increase in spending or decrease in

[2]Japan's gross debt is much larger—roughly 250 percent of GDP—but is offset by significant holdings of assets. For the United States, gross debt for all levels of government was 105 percent, offset by government financial assets (mostly in the state and local sector) equal to 25 percent of GDP.

taxes to spur economic activity, the resulting rise in debt from a level that is already unusually high might cause long-term interest rates to move up sharply, which would limit the effectiveness of such a policy. The surge in federal debt from 35 percent of GDP in 2007 to 74 percent in 2015, and the fact that some analysts wanted fiscal policy to be even more expansionary during those years, illustrates why fiscal space is desirable. As another example, if interest rates increased significantly or trend growth deteriorated significantly, a government with lower debt could make smaller and more gradual changes in taxes and spending than a government already facing higher debt. Auerbach (2016) showed that risk-averse taxpayers would be willing to forego some consumption now to protect themselves against greater reductions in consumption later if the fiscal situation turned out worse than expected.[3]

A further argument for acting promptly to hold down the debt-to-GDP ratio is that doing so would enable policy changes to be made gradually, so that households, businesses, and state and local governments would not be forced to respond suddenly to cope with the associated changes in income and incentives. Following that logic, for example, most proposals that would cut benefits for older Americans do not include cuts for people already receiving benefits or on the cusp of receiving benefits. Indeed, an increase in the full retirement age for Social Security that was enacted in 1983 was phased in so slowly that it will not be fully in place for new beneficiaries until 2022.

Implications of Population Aging

Population aging affects the growth of federal debt for any given budget policies, and it also has implications for the *optimal* amount of debt and thus for *optimal* policies. Our framework for analysis is a standard Solow-style model of economic growth with a society that aims to maximize the present discounted value of its well-being into the indefinite future (for a description of the Solow model used by the Congressional Budget Office, see CBO 2014b). This approach follows that in Cutler, Poterba, Sheiner, and Summers (1990) and Elmendorf and Sheiner (2000); a more detailed analysis underlying the conclusions described here appears in Elmendorf and Sheiner (2016). We begin by considering the macroeconomic implications of aging and then turn to the budgetary implications.

[3] Uncertainty about future interest rates and economic growth rates appears to be the primary reason that US Treasury interest rates have been less than the growth rate of the economy for long historical periods. Ball, Elmendorf, and Mankiw (1998) argued that the marginal product of capital has exceeded, on average, the growth rate of the economy—that is, the economy has been dynamically efficient—but the risk premium has been large enough to push Treasury rates below the growth rate. In that circumstance, there is a "deficit gamble" available to society: Letting debt run up for a period and then restoring primary budget balance would be good for current generations and *probably* not hurt future generations, because debt would *probably* decline again relative to GDP without further policy action. However, that gamble might fail through an increase in interest rates or faltering of growth rates. Moreover, because future policy changes would tend to be needed when growth was low, the policy changes would be especially costly in terms of social welfare. Therefore, Ball, Elmendorf, and Mankiw concluded, the expected value of this gamble from the standpoint of future generations was negative.

Macroeconomics of Population Aging

Population aging in the United States is attributable to two factors: a drop in fertility after the "baby boom" that followed World War II and continued increases in longevity. Both factors are reducing the number of workers relative to the total population, which implies a decline in per capita GDP for any given amount of capital, productivity, and labor force participation. Lower fertility (but not increased longevity) has also reduced the growth of the labor force, which has reduced the saving required to equip new workers with any given amount of capital. Together, those two factors will reduce sustainable per capita consumption a few decades from now by roughly 11 percent relative to what it would be if the population were not aging. Consumption will still be higher in the future, just not as high as it would be if the age distribution of the population was unchanged.

Society could respond to this gradual reduction in sustainable consumption (relative to what would happen in the absence of aging) in various ways. Society could simply allow actual consumption to fall in line with sustainable consumption, which would leave current consumption unaffected but cause future consumption to decline by 11 percent. However, society could instead decrease current consumption, and increase saving and investment, in order to build up its capital stock—which would allow future consumption to decline by less than 11 percent.

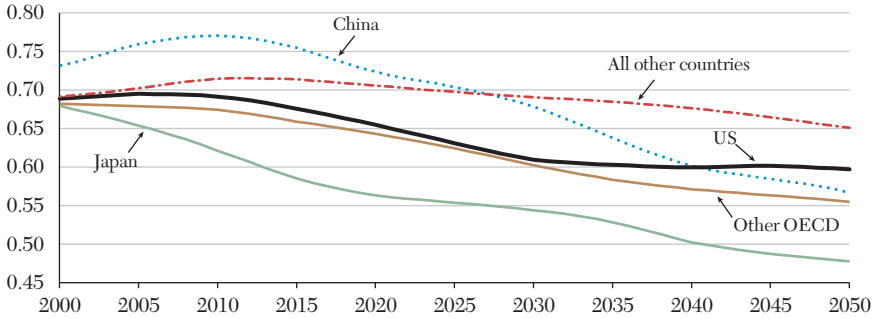
Using a plausibly calibrated growth model and welfare function for society, we estimate that the United States should build up its capital stock over the next decade or two, relative to what would be optimal in the absence of population aging, and then decrease the capital stock later to buffer the decline in consumption.[4] However, the optimal increase in saving turns out to be relatively small: Optimal consumption falls by 4 percent initially and 9 percent over the next two decades, and the maximum increase in the capital–labor ratio is just 6 percent. Optimal saving over the coming decade is higher by roughly 1½ percent of GDP.

Those estimates are based on modeling the United States as a closed economy, but estimates based on an open-economy model are (perhaps surprisingly) similar. We use the same calibrated growth model and welfare function for the United States and add a comparable growth model and welfare function for the rest of the world, allowing for free capital flows between countries (an admittedly extreme assumption).

Based on demographic projections from the World Bank, the ratio of workers to population will decline considerably over the coming decades in each major segment of the world economy outside this country, as shown in Figure 2. Relative to the pattern in the United States, the ratio of workers to population in the rest of the world is expected to fall less sharply in the next few decades and more sharply thereafter, in part because labor force growth will fall even lower in other countries than in the United States. With aging in the rest of the world proceeding at roughly the same rate as aging in the United States, the results of our modeling that includes

[4] That finding differs from the conclusions in Cutler, Poterba, Sheiner, and Summers (1990) and Elmendorf and Sheiner (2000), which—using the same approach—found that the optimal response in 1990 and 2000 to *future* population aging was to *reduce* saving and increase consumption. However, now that the demographic transition has begun, the optimal response in 2016 is to raise saving and decrease consumption.

Figure 2
Ratio of Workers to Population
(ratio of those 15–64 to total population)



Source: World Bank (demographic inputs); National Transfer Accounts and Census (consumption weights for support ratios); authors' calculations.
Note: "Other OECD" includes all countries in the OECD other than the United States and Japan. "All other countries" includes all the countries in the world other than China and the OECD. Support ratios for "Other OECD" and "All other countries" calculated using 2013 GDP-per-worker weights.

the rest of the world are close to our results for the United States alone: Optimal consumption falls by a little over 3 percent initially and 8 percent over the next two decades, and the maximum increase in the capital–labor ratio is 3 percent. Optimal saving over the coming decade is higher by roughly ¾ percent of GDP.

Budgetary Implications of Population Aging

To achieve the desired rise in national saving in response to population aging, one natural approach is to increase federal saving (or decrease federal dissaving). If private saving did not respond to changes in the federal budget, the optimal increase in national saving could be achieved by reducing the federal deficit by 1 percent of GDP—an amount that is about one-twentieth of projected federal spending or revenue and about one-quarter of projected deficits over the coming decade. However, private saving *would* respond to such budgetary changes, because spending cuts or tax increases would affect income and interest rates in ways that would tend to decrease private saving. Moreover, population aging might directly lead to an increase in private saving (as people increase saving in response to longer life expectancy), and population aging might lead to an increase in saving by state and local governments (through responses to projected growth in Medicaid expenditures or looming obligations for retiree pensions or health care, as discussed by Novy-Marx and Rauh 2014 and Lutz and Sheiner 2014). Thus, the implied deficit reduction could be larger or smaller than the optimal change in national saving.

Our analysis of population aging does not include some potentially important considerations. First, our modeling incorporates no change in labor force participation at given ages as the population grows older. Sheiner (2014) calculates that a gradual increase in labor force participation at given ages cumulating to 11 percent

Figure 3
Yield on Treasury Securities

Source: Federal Reserve.

would fully offset the effects of aging on sustainable per capita consumption, and at least some increase in participation at given ages seems likely as life expectancy rises. Second, we have not discussed the benefits of smoothing tax rates over time to minimize deadweight losses. This consideration increases the desirability of raising taxes sooner rather than later (unless deficit reduction would be achieved solely through cuts in projected spending). However, in Elmendorf and Sheiner (2016), we show that this smoothing consideration is not quantitatively important for the issue at hand. Third, our modeling assumes that resources are fully employed at all times, that is the economy never has recessions. But in the real world, in changing fiscal policy to achieve long-term goals, one should make gradual changes rather than sharply increasing saving, which might inadvertently cause a recession.

Implications of Persistently Low Interest Rates on Federal Debt

Interest rates on both short-term and long-term federal debt are now very low by historical standards despite the continuing economic expansion, the onset of tightening in monetary policy by the Federal Reserve, expectations of fiscal expansion under the new president, and the surge in outstanding federal debt since 2007. As illustrated in Figure 3, Treasury yields rose dramatically between the mid-1960s and early 1980s as inflation climbed, and then reversed course equally dramatically as inflation fell. However, even with inflation fairly stable in the 1990s and 2000s before the financial crisis, yields on federal debt continued to fall significantly. Unsurprisingly, yields fell notably further during the crisis and severe recession that followed, as the Federal Reserve cut short-term interest rates and investors sought a safe haven in turbulent markets. More surprisingly, yields have rebounded only to a limited extent in the past several years in spite of the factors just mentioned.

To explain low rates and assess their likely persistence, a number of researchers at institutions like the Federal Reserve, the International Monetary Fund, and the Bank of England, as well as noted economists like Larry Summers, Ben Bernanke, and Paul Krugman have reviewed or attempted to quantify the impact on interest rates of a wide range of factors.[5] Those analyses have generally concluded that interest rates will increase over the next several years but remain significantly below their average levels of the past few decades. For example, the Congressional Budget Office (2016) projects that the yield on 10-year Treasury notes will average 4.3 percent over the next 30 years, compared with 5.8 percent during the 1990–2007 period of low inflation and fairly stable economic and financial conditions. Also, the median forecast of the federal funds rate "in the longer run" by members of the Federal Open Market Committee (FOMC) is 3 percent, compared with a 4.4 percent average during the 1990–2007 period (Federal Reserve 2016). Interest rates in other countries are also expected to be significantly lower in coming years than in the past. The International Monetary Fund (2016) projects that the real long-term interest rate on government securities—represented by a weighted average of rates on 10-year securities from different countries—will be 0.2 percent at the end of this decade, compared to a 1998–2007 average of 2.4 percent.

To be sure, the outlook for interest rates on federal debt is highly uncertain: Projecting factors that affect interest rates and quantifying their influence on rates is difficult, and financial market participants and economic forecasters may have overreacted to the experience since the financial crisis. Federal budget policy should allow for the risk that rates rise substantially in the years ahead. But both market prices and published analyses imply that a more likely outcome is low rates for an extended period.

We turn to documenting the implications of persistently low Treasury interest rates for federal debt dynamics. Then we examine the implications of low Treasury interest rates for optimal federal debt in two cases: when rates are low because the marginal product of private capital is low; and when those rates have fallen relative to the marginal product of private capital. We further explore the implications of low interest rates for countercyclical federal budget policy and for federal investment.

Implications for Federal Debt Dynamics of Persistently Low Interest Rates

For any given paths of federal revenues and noninterest spending, persistently low interest rates reduce future debt. Between 2013 and 2016, the Congressional Budget Office revised downwards its projection of average 10-year Treasury note rates over the following 30 years by about 1 percentage point. Moreover, CBO (2016) estimated that, if interest rates on federal debt were 1 percentage point lower than the agency expects during the next 30 years, federal debt would be smaller by more than 30 percent of GDP at the end of that period.

[5]Examples include Bean, Broda, Ito, and Kroszner (2015), Bernanke (2015a, b, c, d), CBO (2014a), Council of Economic Advisers (2015), Federal Reserve Board (2015), Hamilton, Harris, Hatzius, and West (2015), International Monetary Fund (2014), Krugman (2015), Rachel and Smith (2015), Summers (2013a, 2014, 2015a, b), and Thwaites (2015).

However, the dynamics of federal debt are also affected significantly by the rate of economic growth, for which projections have also been revised downwards over the past few years. With the baby boom generation heading into retirement and the labor force participation rate among working-age women roughly stabilizing after increasing sharply for a few decades, the US labor force will grow much more slowly in the next few decades than in the past few decades. Moreover, disappointing productivity growth in recent years has led the Congressional Budget Office and other analysts to lower their expectations for future productivity gains. Between 2013 and 2016, CBO's downward revision to projected growth during the next 30 years raised the projected ratio of debt-to-GDP 30 years ahead by roughly 15 percentage points.

Implications for Optimal Federal Debt of a Lower Marginal Product of Private Capital

One of the reasons that interest rates will probably be lower in the next few decades than in previous decades is that the marginal product of private capital—the return to additional private investment—will probably be lower. Many different factors may play a role in reducing the marginal product of capital, and they have different implications for budget policy.[6]

First, the marginal product of capital may decline because of population aging. Slower growth in the labor force as the baby boom generation retires, if not accompanied by a corresponding reduction in investment, raises the amount of capital per worker and pushes down its marginal return. As we explained above, the government should respond to aging by decreasing current consumption and *increasing* current national saving and investment (causing the capital–labor ratio to rise), which it can accomplish by issuing less debt than otherwise.

Second, the marginal product of private capital may be held down by slower economic growth stemming from slower growth in total factor productivity.[7] Slower productivity growth diminishes the return to additional national saving—in other words, it raises the price of future national consumption relative to current consumption—which implies that the government should decrease saving. However, slower productivity growth also means that future generations will not be as much better off relative to current generations as they would be otherwise, which implies that the government should increase saving. Using the same growth model and welfare function that we applied to aging, we estimate that, on balance, the government should respond to lower productivity growth by *slightly* increasing national saving, which it can accomplish by issuing slightly less debt than otherwise.

[6] For more discussion of the factors listed in the following paragraphs, see Bernanke (2005), Caballero, Farhi, and Gourinchas (2008, 2015), CBO (2016a), Dynan, Skinner, and Zeldes (2004), Furman (2015b), Mericle and Struven (2016), and Summers (2014, 2015a, b).

[7] As noted earlier, economic growth will also be slower than in the past because of a leveling off in women's labor force participation after sharp increases in the 1970s through 1990s. To the extent that the previous increases reflected increasing opportunities for women in the labor force or shifts in social norms, the implications for federal budget policy of the leveling off in participation are similar to the implications of slower productivity growth.

Third, the marginal product of capital may be held down by an increase in private saving that raises the amount of capital per worker. For households, increasing income inequality pushes up personal saving because of the greater saving propensity of higher-income people; for businesses, a high level of profits relative to national income is supporting business saving. On the other hand, the downward trend of the personal saving rate during the past few decades suggests that factors other than increasing inequality have had larger effects on private saving, and in coming years the retirement of the baby boomers will shift more people from their years of peak saving to years of lower saving or dissaving. If Americans choose to save more privately because of a shift in their preferences, the government should *not* try to undo that choice by issuing additional debt, and perhaps the government should accommodate the shift in preferences by issuing less debt than otherwise.

Fourth, the marginal product of private capital will probably be held down by increased capital inflows. Slow growth in many economies around the world and consequent declines in the return on investment in those countries appear to be sustaining a so-called global savings glut—even though some observers have expected that emerging market economies, at least, would choose to invest more at home in response to strong domestic investment opportunities and an already high level of overseas assets. With a savings glut, the return to additional US national saving is lower and the government should decrease saving, which it can accomplish by issuing more debt than otherwise.

Fifth, the marginal product of private capital may be held down by a decline in the capital intensity of production arising from the growing importance of sectors that use little physical capital and by a continuing drop in the price of information technology that allows any given amount of inflation-adjusted investment to be achieved with a smaller amount of nominal investment. Those changes diminish the demand for capital, which means that the return to additional national saving is lower. As a consequence, the government should decrease national saving, which it can accomplish by issuing more debt than otherwise.

How should one weigh these various factors that may be reducing the marginal product of capital? The first (population aging) was covered in our earlier discussion of the implications of aging; the second (slower economic growth stemming from slower growth in total factor productivity) looks quantitatively unimportant in our modeling; the third (an increase in private saving) is probably not an important part of the explanation for low interest rates given that private saving has decreased over the past decade; and the fourth and fifth explanations (increased capital inflows, a decline in the capital intensity of production) suggest that the government should issue additional debt. Therefore, we conclude that the decline in the marginal product of private capital apart from the effect of aging implies that the government should decrease national saving relative to what it would be otherwise, which the government can accomplish by issuing more debt than it would otherwise.

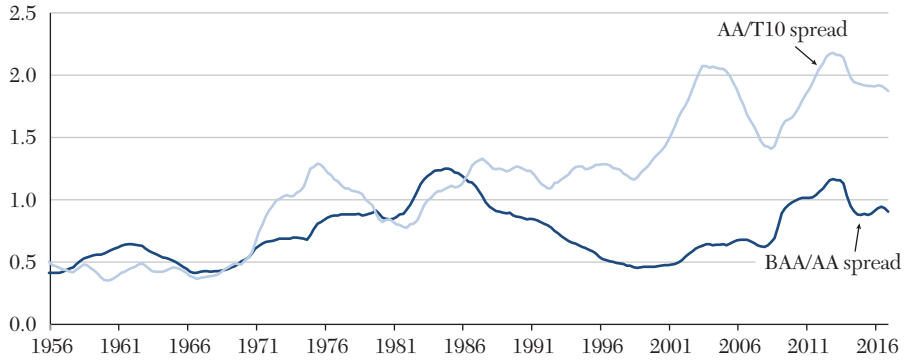
If the marginal product of private capital will be lower than in the past, the marginal product of public capital probably will be lower as well, but not as much

lower. Among the factors that will reduce the return to private capital, some will also reduce the return to public capital (for example, slower labor force growth diminishes the value of additional spending for education in the same way as it diminishes the value of additional business equipment), but others will not (for example, inflows of foreign capital to businesses and increases in private saving in response to aging do not increase public capital). Therefore, the return to public capital relative to the return to private capital will probably increase, so federal investment in physical and human capital should increase. We return to this issue below.

Recent patterns of business investment and capital income do not provide clear evidence that the marginal product of private capital has already declined. On the one hand, a drop in the marginal product of physical capital could explain why investment has not increased markedly in the past few years despite the very low cost of funds and presumably some pent-up need for capital following weak investment during the downturn—but on the other hand, investment may also be held down by lingering concerns about the strength of demand for goods and services, uncertainty about future tax policy, or other factors. In addition, measured investment during the past several years does not imply an increase in capital per worker of the sort implied by some of the explanations for a declining marginal product. Capital income is now a historically high share of national income, but that fact does not have clear implications for the marginal product of capital: the relationship between the capital share and the marginal product depends on the production function, and the capital share may be high because of factors that are not captured in standard production functions and that have different possible implications for the marginal product. For example, measured capital income probably includes some returns to human and intangible capital, which may have increased in relative importance over time; firms may be receiving greater monopoly rents; globalization or changes in social norms may be allowing firms to capture a larger share of total product; and nondiversifiable investment risk may have increased, with the average return on capital rising to compensate.

One possibility that is related but not identical to the possibility of a declining marginal product is that desired saving in this country has increased in recent years but investment demand is so insensitive to the cost of capital that the quantity of investment has not increased much and instead the return on assets has been pushed down. A small response of investment to an increase in desired saving, and consequent downward pressure on returns, would be consistent with some of the empirical literature (for example, Tevlin and Whelan 2003; Kothari, Lewellen, and Warner 2014; Pinto and Tevlin 2014; Banerjee, Kearns, and Lombardi 2015) and with evidence that firms' hurdle rates for returns on new investments tend to be insensitive to the cost of capital and have not fallen in the past several years (Sharpe and Suarez 2014)—but not consistent with other parts of the empirical literature (for example, see Cummins, Hassett, and Hubbard 1994; Gilchrist and Zakrajsek 2007). If desired saving has increased but the quantity of investment has not increased much, the result would be a drop in interest rates on private bonds, an increase in price-earnings ratios on equities, low business interest

Figure 4
BAA to AA and AA to 10-year Treasury Spreads
(percentage points; five-year moving average)



Source: Bloomberg; Federal Reserve.
Note: "T10" means 10-year Treasury notes. The widening spread between the yields on 10-year Treasury notes and BAA-rated corporate debt since 1980 can be attributed entirely to an increase in the spread between the yield on AA-rated debt and the yield on Treasury debt with little change, on net, in the spread between the BAA and AA yields.

payments, and high profits—all of which are true today. The implications of this scenario for federal budget policy are the same as the implications of a declining marginal product caused by increased capital inflows: because an increase in desired saving is not finding a productive use through private investment, both federal debt and federal investment in physical and human capital should be higher than otherwise.

Implications for Optimal Federal Debt of a Larger Difference Between the Marginal Product of Private Capital and the Yield on Treasury Debt

Interest rates on federal debt also may be lower than in the past because those rates have fallen relative to the marginal product of private capital. One possibility is that the perceived risk of private capital has increased or the price of that risk has increased. The financial crisis, severe recession, and slow recovery certainly represent a stark change from the so-called "Great Moderation" of the economy between the mid-1980s and mid-2000s. But that story cannot explain the downtrend in yields on US Treasury debt before the financial crisis, nor is it consistent with the fact that the widening spread between the yields on 10-year Treasury notes and BAA-rated corporate debt since 1980 can be attributed entirely to an increase in the spread between the yield on AA-rated debt and the yield on Treasury debt with little change in the spread between the BAA and AA yields, as shown in Figure 4.

If interest rates on federal debt are lower because people perceive that the risk of private capital or the price of that risk have increased, the price of future national consumption relative to current consumption has not changed on a risk-adjusted basis. Therefore, the federal government should not try to change national

saving by changing the amount of debt it issues.[8] The larger difference between the average returns on private securities and Treasury securities would increase the average gain for the government of issuing debt in order to purchase private securities, but because that larger difference reflects greater risk or a higher price on risk in this scenario, the government should not follow this strategy unless the federal government's ability to bear risk, relative to that of the private sector, has increased as well. In addition, the risk-adjusted return to public capital relative to private capital is unchanged in this scenario, so federal investment should not change.

A different possibility is that factors besides changing assessments of risk or the price of risk may have increased the demand for federal debt. Global GDP has been increasing more rapidly than US GDP, so total foreign demand for the safety and liquidity of Treasury securities has probably increased significantly even apart from any reassessment of risk. In addition, financial regulations now require certain institutions to maintain greater amounts of capital and liquidity than before the financial crisis, and federal debt is valuable for satisfying those requirements. Of course, the supply of Treasury securities has more than doubled in the past eight years, which one might expect to have offset any increase in demand for those reasons. But some analysts have argued that the supply of assets *perceived* as safe has actually fallen (Caballero and Farhi 2014).

If interest rates on federal debt are lower because factors other than changes in the perceived risk, or the price of risk, of private assets have increased the demand for federal debt, the implications for federal debt are subtler. The government's ability to borrow more cheaply from domestic investors—who currently hold about half of federal debt—represents an implicit tax on those investors who own federal debt. This phenomenon is sometimes termed "financial repression" (as in Reinhart and Sbrancia 2011). The government's ability to borrow more cheaply from foreign investors—who currently hold roughly the other half of federal debt—represents an opportunity to extract resources from nonresidents at lower cost than otherwise. Increasing federal debt is the optimal response when considering both groups, because the resulting higher interest rates would return the implicit tax on domestic investors toward its previous level (which would be appropriate if the previous level was chosen optimally) and because the greater amount of borrowing would increase the extraction of resources from foreign investors (which makes sense because the cost of extraction is lower than it was previously).

However, in this scenario, there is no change in the return on private assets and therefore no change in the price of future national consumption relative to current consumption—which means that the federal government should not try to change national saving. Moreover, the risk-adjusted return to public capital relative to private capital is unchanged, so federal investment in physical and human

[8] That conclusion would not necessarily hold, though, if an increase in government debt would reduce the risk premium on private assets, perhaps by diminishing fears of secular stagnation. We discuss the issue of secular stagnation later.

capital should not change. What, then, should the additional funds from issuing more federal debt be used for?

If other considerations do not disallow as inappropriate, the funds should be used for federal purchases of private financial assets, because then neither national saving nor public investment would be altered; in addition, the resulting increase in risk-taking on the federal government's balance sheet would be appropriate because the spread between the returns on private securities and Treasury securities is greater in this scenario without any change in perceived risk or its price. Yet federal purchases of private financial assets may not be appropriate, perhaps because of the difficulty of the government's purchasing and holding assets neutrally across companies and sectors. In that case, the additional funds raised by issuing more debt should be used for higher federal spending (both nonfinancial investment and consumption) and lower taxes, although the distortions created from these policies mean that less additional debt should be issued than if private financial assets were being purchased.

Implications of Persistently Low Interest Rates for Countercyclical Federal Budget Policy

Persistently low interest rates on federal debt—for whatever reason—will limit the ability of monetary policy to counteract future recessions. In each of the past three economic downturns, the Federal Reserve has cut the federal funds rate by more than 5 percentage points. However, the federal funds rate is highly unlikely to reach 5 percent for the foreseeable future, so when the next recession occurs, the Federal Reserve will be unable to ease monetary policy nearly as much as it has in the past.[9]

How should federal budget policy respond? Part of the answer is that federal debt should be higher, on average. Additional outstanding debt would tend to raise interest rates, which would give the Federal Reserve more room to cut the funds rate when recessions hit. Moreover, additional debt would increase the amount of government securities the Federal Reserve could purchase to achieve quantitative easing.

Another part of the answer is that federal debt should vary more over the business cycle. To achieve that greater variation, policymakers should build automatic fiscal stabilizers that are more powerful than the existing stabilizers and that respond rapidly to changing conditions, and policymakers should enact further spending increases and tax cuts when conditions warrant.

Suppose that interest rates were not just persistently low and sometimes close to zero, but instead were stuck close to zero on an ongoing basis—or, in other words, suppose that the Federal Reserve was not just constrained from reducing the federal funds rate enough to achieve full employment periodically, but was constrained

[9]The Federal Reserve has other potential ways to strengthen its response to future recessions. For example, it could do more quantitative easing, lower the federal funds rate below zero, or raise inflation (and thus the funds rate) before the next downturn. But many observers think those approaches would be less potent than traditional monetary policy or present difficulties of their own. See, for example, Cúrdia and Ferrero (2013) and Yellen (2015).

continually. In that setting, aggregate demand would be so weak that output would fall below its potential even with a funds rate near zero.

Summers (2013b, 2014, 2015a, b) and others have deemed such a situation as "secular stagnation" and have argued that it can occur because of weakness in either domestic demand or foreign demand for US goods and services (see also Bernanke 2015a, b, c, d; Krugman 2015; Teulings and Baldwin 2014). For example, if foreigners' demand for US assets increased because of stagnation in other countries, the resulting inflow of funds would push up the exchange value of the dollar and reduce US net exports; in the words of Caballero, Farhi, and Gourinchas (2015), when interest rates are extremely low around the world, "lower global output … rebalances global asset markets … [as] liquidity traps emerge naturally and countries drag each other into them." Secular stagnation also can be self-reinforcing within a country, because weakness in output relative to potential tends to reduce inflation, which raises inflation-adjusted interest rates if nominal rates are stuck near zero. Whether the US economy will experience secular stagnation in coming years is unclear. Although economic growth has been tepid in the past few years, the unemployment rate has declined considerably, implying that a federal funds rate just above zero, a considerable amount of quantitative easing, and ongoing federal deficits have caused actual output to increase more rapidly than potential output. Still, with interest rates expected to be low for years to come—in this country and others—secular stagnation is clearly a risk.

If interest rates were stuck close to zero on an ongoing basis, federal debt should be higher than otherwise. However, that situation would not continue indefinitely: rising federal debt would ultimately tend to increase interest rates, in part by raising the perceived riskiness of that debt.

Implications of Persistently Low Interest Rates for Federal Investment

As we explained above, many of the reasons for persistently low interest rates on federal debt imply that federal investment in physical and human capital should be higher than would otherwise be optimal. Assessing whether such investment should be increased *from current levels* requires a broader assessment of the marginal costs and benefits of that investment, which lies beyond the scope of this paper. However, under the current limits on annual appropriations, federal nondefense investment as classified by the Office of Management and Budget (2016)—which includes infrastructure investment, research and development, and some support for education and training—will soon fall to the smallest percentage of GDP in at least half a century. Therefore, just maintaining the historical levels of such investments would require a significant increase relative to what will occur under current law.

Generally, the federal government should undertake all investments for which the risk-adjusted social return is greater than the social cost of the required resources. The social return of an investment includes the increment to GDP arising from the investment as well as benefits that are not measured in GDP such as better air quality or a longer life expectancy. The social cost of an investment depends on the value of private investment that is crowded out and the

deadweight loss from the distortionary taxation needed to finance it. This criterion is different from whether a public investment would raise GDP by enough to "pay for itself" through extra tax revenue because that criterion counts on the benefit side only the tax revenue generated as a result of the investment, and counts on the cost side the budgetary cost rather than the value of lost private investment and the deadweight loss of financing. (For contending views on whether public investments pay for themselves in a budgetary sense, see DeLong and Summers 2012, Congressional Budget Office 2016, and Summers 2016.) However, the more tax revenue that a federal investment generates, the less distortionary taxation is required to finance it and, thus, the lower is the social cost (all else equal).

The question of how an investment should be financed—through borrowing, or through higher taxes or lower spending of other sorts—is separate from whether the investment should be undertaken. How best to finance an investment depends on whether the federal government is trying to increase national saving at the time the investment is made. Federal investment that is financed by borrowing leaves public saving unchanged (because the extra saving in the form of the investment offsets the dissaving seen in the larger budget deficit) and probably raises private saving a little (because the additional federal borrowing raises interest rates)—thereby raising national saving a little. Federal investment that is financed by tax increases or spending cuts increases national saving more notably because public saving increases (through the investment) and private saving is essentially unchanged (because changes in federal spending have little direct effect on private consumption and changes in taxes induce roughly corresponding changes in private consumption in the long run). Therefore, if the government believes that national saving is already optimal, it should finance worthwhile investments (those with a social return greater than the social cost) through borrowing, while if the government wants to increase national saving, it should finance those investments by raising taxes or cutting other spending. We emphasize, however, that the decision about whether to undertake a public investment should depend only on the net social return and not on the means of financing, which should be decided separately based on the optimal amount of national saving. Similarly, concerns about fiscal space should affect how federal investments are financed but should not affect whether specific investments are made—except in cases where federal investment boosts tax revenue sufficiently to increase fiscal space.

The social return to federal investment is difficult to assess and likely varies significantly across investments. Returns to highways, for example, have been the subject of research for decades, while returns to many other types of investments have not. Moreover, returns vary considerably within categories; improving key highway links has a higher return than building "bridges to nowhere." And some types of federal spending not traditionally classified as investment have an element of investment. For example, certain benefits for lower-income families have been shown to increase the future earnings of their children (for an overview of recent research on such investments, see Furman 2015a and Butcher 2017). Improving the selection of federal investments through more rigorous analysis could increase the average return.

Conclusion

The ratio of federal debt to GDP will almost surely continue to rise unless the country makes significant changes in spending programs, the tax code, or both. Because federal debt is already historically high relative to GDP, and because regaining fiscal space would enhance the government's ability to respond to unexpected events, one might presume that those spending reductions and tax increases should be implemented sooner rather than later. The economic effects of the aging of the US population reinforce the case for reducing deficits quickly.

However, that conventional wisdom overlooks the implications of persistently low interest rates on federal debt: not only do low rates slow the accumulation of debt for given paths of revenues and noninterest spending, they also imply (for many possible explanations of low rates) that both federal debt and federal investment should be higher than they would be otherwise. As a result, the policy changes that will be needed to put federal debt on a sustainable trajectory in the long run should *not* be implemented now, although *enacting* changes now would give households, businesses, and state and local governments time to adjust. For example, a combination of gradual reductions in Social Security and Medicare spending (such as through phased changes in the Social Security benefit formula and in income-related Medicare premiums), increases in taxes, and significantly higher federal investment during the next decade could allow federal debt to rise further relative to GDP over the decade but then to level out or decline relative to GDP in later years. Another implication of persistently low interest rates is that monetary policy will be less effective at responding to recessions and therefore federal budget policy should be more strongly countercyclical than it has been in the past. For example, if laws were changed so that payroll tax rates and the federal share of Medicaid spending depended explicitly on cyclical conditions, automatic fiscal stabilizers would counter future economic downturns more effectively.

■ *We are grateful to Peter Olson for excellent research assistance and to Alan Auerbach, Olivier Blanchard, Greg Duffee, John Fernald, Mark Gertler, Gordon Hanson, Enrico Moretti, Larry Summers, Timothy Taylor, David Weil, and David Wessel for helpful comments.*

References

Auerbach, Alan J. 2016. "Fiscal Uncertainty and How to Deal With It." Working Paper 6, Hutchins Center on Fiscal and Monetary Policy at Brookings, December 15. https://www.brookings.edu/wp-content/uploads/2016/06/15_fiscal_uncertainty_auerbach.pdf.

Ball, Laurence, Douglas W. Elmendorf, and N. Gregory Mankiw. 1998. "The Deficit Gamble." *Journal of Money, Credit, and Banking* 30(4): 699–720.

Banerjee, Ryan, Jonathan Kearns, and Marco Lombardi. 2015. "(Why) Is Investment Weak?" *BIS Quarterly Review*, March, pp. 67–82.

Bean, Charles, Christian Broda, Takatoshi Ito, and Randall Kroszner. 2015. *Low for Long? Causes and Consequences of Persistently Low Interest Rates.* 17th CEPR-ICMB Geneva Report on the World Economy. London: CEPR Press.

Bernanke, Ben S. 2005. "The Global Saving Glut and the U.S. Current Account Deficit." Remarks at the Sandridge Lecture, Virginia Association of Economists, Richmond, Virginia, March 10. Federal Reserve Board. https://www.federalreserve.gov/boarddocs/speeches/2005/200503102/.

Bernanke, Ben S. 2015a. "Why Are Interest Rates So Low." Brookings, March 30. https://www.brookings.edu/blog/ben-bernanke/2015/03/30/why-are-interest-rates-so-low/.

Bernanke, Ben S. 2015b. "Why Are Interest Rates So Low, Part 2: Secular Stagnation." Brookings, March 31. https://www.brookings.edu/blog/ben-bernanke/2015/03/31/why-are-interest-rates-so-low-part-2-secular-stagnation/.

Bernanke, Ben S. 2015c. "Why Are Interest Rates So Low, Part 3: The Global Savings Glut." April 1. https://www.brookings.edu/blog/ben-bernanke/2015/04/01/why-are-interest-rates-so-low-part-3-the-global-savings-glut/.

Bernanke, Ben. 2015d. "Why Are Interest Rates So Low, Part 4: Term Premiums." Brookings, April, 13. https://www.brookings.edu/blog/ben-bernanke/2015/04/13/why-are-interest-rates-so-low-part-4-term-premiums/.

Butcher, Kristin. 2017. "Assessing the Long-Run Benefits of Transfers to Low-Income Families." Paper prepared for the Hutchins Center on Fiscal and Monetary Policy at Brookings' conference "From Bridges to Education: Best Bets for Public Investment," held at Brookings, January 9, 2017. https://www.brookings.edu/research/assessing-the-long-run-benefits-of-transfers-to-low-income-families/.

Caballero, Ricardo J., and Emmanuel Farhi. 2014. "On the Role of Safe Asset Shortages in Secular Stagnation." Chap. 9 in *Secular Stagnation: Facts, Causes and Cures,* edited by Coen Teulings and Richard Baldwin. CEPR Press.

Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2008. "An Equilibrium Model of 'Global Imbalances' and Low Interest Rates." *American Economic Review* 98(1): 358–93.

Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2015. "Global Imbalances and Currency Wars at the ZLB." NBER Working Paper 21670, October.

Congressional Budget Office (CBO). 2010. "Federal Debt and Interest Costs." December 14.

Congressional Budget Office (CBO). 2014a. *The 2014 Long-Term Budget Outlook.* July 15.

Congressional Budget Office (CBO). 2014b. "How CBO Analyzes the Effects of Changes in Federal Fiscal Policies on the Economy." November 10.

Congressional Budget Office (CBO). 2016a. *The Budget and Economic Outlook: 2016 to 2026.* January 25.

Congressional Budget Office (CBO). 2016b. *The 2016 Long-Term Budget Outlook.* July 12.

Council of Economic Advisers. 2015. "Long-Term Interest Rates: A Survey." July.

Cummins, Jason G., Kevin A. Hassett, and R. Glenn Hubbard. 1994. "A Reconsideration of Investment Behavior Using Tax Reforms as Natural Experiments." *Brookings Papers on Economic Activity,* no. 2, pp. 1–74.

Cúrdia, Vasco, and Andrea Ferrero. 2013. "How Stimulatory Are Large-Scale Asset Purchases?" FRBSF Economic Letter 2013-22, August.

Cutler, David M., James M. Poterba, Louise M. Sheiner, and Lawrence H. Summers. 1990. "An Aging Society: Opportunity or Challenge?" *Brookings Papers on Economic Activity,* no. 1, pp. 1–73.

DeLong, J. Bradford, and Lawrence H. Summers. 2012. "Fiscal Policy in a Depressed Economy." *Brookings Papers on Economic Activity,* Spring, pp. 233–97.

Dynan, Karen E., Jonathan Skinner, and Stephen P. Zeldes. 2004. "Do the Rich Save More?" *Journal of Political Economy* 112(2): 397–444.

Elmendorf, Douglas W., and Louise M. Sheiner. 2000. "Should America Save for Its Old Age? Fiscal Policy, Population Aging, and National Saving." *Journal of Economic Perspectives* 14(3): 57–74.

Elmendorf, Douglas, and Louise Sheiner. 2016. "Federal Budget Policy with an Aging Population and Persistently Low Interest Rates." Hutchins Center Working Paper 18, Hutchins Center on Fiscal and Monetary Policy, October.

Federal Reserve Board. 2015. "Minutes of the Federal Open Market Committee, October 27–28."

Federal Reserve Board. 2016. "Economic

Projections of Federal Reserve Board Members and Federal Reserve Bank Presidents." December. https://www.federalreserve.gov/monetarypolicy/files/fomcprojtabl20161214.pdf.

Furman, Jason. 2015a. "Smart Social Programs." *New York Times*, Opinion Pages, May 11. http://www.nytimes.com/2015/05/11/opinion/smart-social-programs.html.

Furman, Jason. 2015b. "Business Investment in the United States: Facts, Explanations, Puzzles, and Policies." Unpublished paper, September.

Gilchrist, Simon, and Egon Zakrajsek. 2007. "Investment and the Cost of Capital: New Evidence from the Corporate Bond Market." NBER Working Paper 13174, June.

Hamilton, James D., Ethan S. Harris, Jan Hatzius, and Kenneth D. West. 2015. "The Equilibrium Real Funds Rate: Past, Present, and Future." Working Paper 16, Hutchins Center on Fiscal & Monetary Policy at Brookings. October, 30.

International Monetary Fund. 2014. "Perspectives on Global Interest Rates." Chapt. 3 in World Economic Outlook: Recovery Strengthens, Remains Uneven. World Economic and Financial Surveys, April 2014. International Monetary Fund.

International Monetary Fund. 2016. World Economic Outlook: Subdued Demand: Symptoms and Remedies. World Economic and Financial Surveys, October 2016.

Kothari, S. P., Jonathan Lewellen, and Jerold B. Warner. 2014. "The Behavior of Aggregate Corporate Investment. Simon Business School Working Paper no. FR 14-18. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2511268.

Krugman, Paul. 2015. "Liquidity Traps: Local and Global." New York Times, Opinion Pages, April, 1.

Lutz, Byron, and Louise Sheiner. 2014. "The Fiscal Stress Arising from State and Local Retiree Health Obligations." Journal of Health Economics 38: 130–46. December.

Mericle, David, and Daan Struyven. 2016. "US Daily: Capital-Light Technologies and the Long-Term Capex Outlook." Goldman Sachs Economic Research, January 20.

Novy-Marz, Robert, and Joshua Rauh. 2014. "Revenue Demands of Public Employee Pension Promises." American Economic Journal: Economic Policy 6(1): 193–229.

Office of Management and Budget. 2016. Analytical Perspectives: Budget of the United Sates Government, Fiscal Year 2017. February.

Pinto, Eugenio, and Stacey Tevlin. 2014. "Perspectives on the Recent Weakness in Investment." Board of Governors of the Federal Reserve System. *FEDS Notes*, May 21.

Rachel, Lukasz, and Thomas D. Smith. 2015. "Secular Drivers of the Global Real Interest Rate." Bank of England Staff Working Paper 571. December.

Reinhart, Carmen M., and M. Belen Sbrancia. 2011. "The Liquidation of Government Debt." NBER Working Paper 16893.

Sharpe, Steve A., and Gustavo A. Suarez. 2014. "The Insensitivity of Investment to Interest Rates: Evidence from a Survey of CFOs." FEDS Working Paper 2014-02, The Federal Reserve Board.

Sheiner, Louise. 2014. "The Determinants of the Macroeconomic Implications of Aging." *American Economic Review* 104(5): 218–23.

Summers, Lawrence H. 2013a. "The Battle over the US Budget is the Wrong Fight." Blog post, October 14. http://larrysummers.com/the-battle-over-the-us-budget-is-the-wrong-fight/.

Summers, Lawrence H. 2013b. "Remarks at International Monetary Fund Economic Forum: Policy Responses to Crises." November.

Summers, Lawrence H. 2014. "U.S. Economic Prospects: Secular Stagnation, Hysteresis, and the Zero Lower Bound." *Business Economics* 49(2): 65–73.

Summers, Lawrence H. 2015a. "On Secular Stagnation: Larry Summers Responds to Ben Bernanke." April.

Summers, Lawrence H. 2015b. "Rethinking Secular Stagnation after Seventeen Months." IMF Rethinking Macro III Conference, April, 16.

Summers, Lawrence H. 2016. "A postscript to Delong and Krugman." January 2. larrysummers.com/2016/01/02/a-postscript-to-delong-and-krugman/.

Teulings, Coen, and Richard Baldwin, eds. 2014. *Secular Stagnation: Facts, Causes and Cures.* CEPR Press.

Tevlin, Stacey, and Karl Whelan. 2003. "Explaining the Investment Boom of the 1990s." *Journal of Money, Credit, and Banking* 35(1): 1–22.

Thwaites, Gregory. 2015. "Why Are Real Interest Rates So Low? Secular Stagnation and the Relative Price of Investment Goods." Bank of England Staff Working Paper 564. November.

World Bank. 2015. "Health, Nutrition and Population Data and Statistics." September. http://datatopics.worldbank.org/hnp/.

Yellen, Janet. 2015. "Inflation Dynamics and Monetary Policy." Speech at the Philip Gamble Memorial Lecture, University of Massachusetts, Amherst. Federal Reserve Board, September 24.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 195–214

How Digitization Has Created a Golden Age of Music, Movies, Books, and Television

Joel Waldfogel

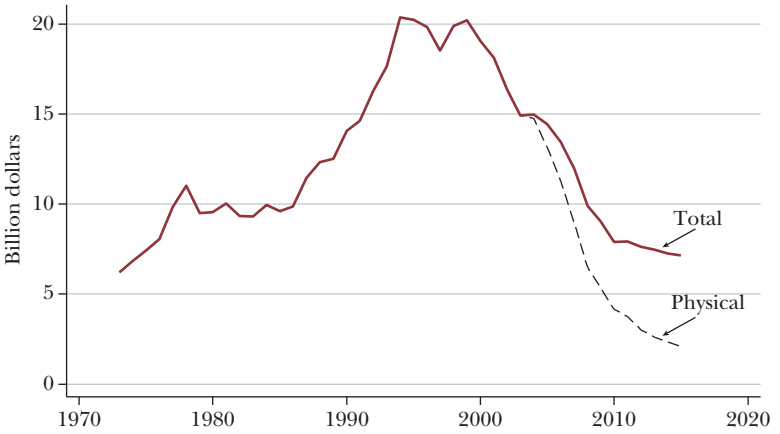
Digitization is disrupting a number of copyright-protected media industries, including books, music, radio, television, and movies. Once information is transformed into digital form, it can be copied and distributed at near-zero marginal costs. This change has facilitated piracy in some industries, which in turn has made it difficult for commercial sellers to continue generating the same levels of revenue for bringing products to market in the traditional ways. The recorded music industry offers a vivid example. Revenue in the recorded music industry had grown steadily throughout the twentieth century but began a precipitous slide in 1999 and has now fallen by more than half (see Figure 1). Yet despite the sharp revenue reductions for recorded music, as well as threats to revenue in some other traditional media industries, other aspects of digitization have had the offsetting effects of reducing the costs of bringing new products to market in music, movies, books, and television. On balance, digitization has increased the number of new products that are created and made available to consumers. Moreover, given the unpredictable nature of product quality, growth in new products has given rise to substantial increases in the quality of the best products and therefore the benefit of these new products to consumers.

■ *Joel Waldfogel is the Frederick R. Kappel Chair in Applied Economics, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is jwaldfog@umn.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.195

doi=10.1257/jep.31.3.195

Figure 1
Total Value of Annual US Music Shipments



Source: Author's calculations from Recording Industry Association of America's reported annual value of music shipments, along with inflation adjustment using the Consumer Price Index.
Note: The figure shows the annual retail value of recorded music sold in the United States, according to the Recording Industry Association of America, in constant 2016 dollars inflated using the Consumer Price Index.

We begin with a discussion of how digitization has threatened the traditional revenue sources for some of these media industries, notably recorded music. We then turn to how digitization has greatly reduced the cost of bringing new products to market in music, books, movies, and television. The reduction in production costs has made the launch of new products in these markets much easier. However, the disruptions and reductions of revenue streams have challenged the roles of the traditional gatekeepers of quality in these industries, including book publishers, recording labels, movie studios, and television networks. These developments have raised concerns that consumer welfare from these media products would fall, on the grounds that high-quality products could not be produced profitably and consumers would be flooded with low-cost and lower-quality products. However, the opposite scenario has emerged—a golden age for consumers who wish to consume media products. For example, consumers have benefited from a wider range of media products available. Moreover, because traditional gatekeepers were imperfect judges of quality, the growth in new products that turn out to be of high quality has presented consumers with many products that would not have met the approval of traditional, pre-digital gatekeepers. Of course, measuring quality in these industries in a direct way is impossible, but a variety of ratings by experts and the public, as well as direct evidence from consumption patterns, suggest that product quality has risen in these media industries. We take up the question of how revenue streams for these industries may be rebuilt in the future, with a focus on the potential of bundled sales strategies and live performance. I conclude with some implications for public policy and future research.

Digitization, Round 1: Piracy and Revenue Reduction in Music

Media industries have several main sources of revenue, including advertising as well as direct and indirect payments from users. Digitization has disrupted many of these streams.

In the recorded music industry, the first manifestations of digitization were sharp and sustained reductions in revenue, and the proximate cause was almost surely piracy. With the appearance of the Napster peer-to-peer file sharing service in 1999, consumers could access almost any piece of recorded music without making a payment to a rights-holder. Recorded music revenue began a sharp and sustained slide, both in the United States and around the world, in 1999. As Figure 1 shows, the rise in digital music sales after 1999 did not nearly offset the decline in physical sales. In the early 2000s, scholars debated whether unpaid access to recorded music would stimulate or depress demand for paid access over time. Perhaps taste-setting consumers involved in music file-sharing would sift through new music, making suggestions to follow-on consumers, and thus stimulate demand for purchasing additional music? After all, when radio broadcasting arrived in 1920s, sales of recorded music fell for a time, but then recorded music revenue rose fairly steadily after the 1920s (Liebowitz 2004).

The task of estimating how the growth in web-based music file-sharing depressed revenue turned out to be difficult to address, for a variety of reasons. While data on legal sales are available, data on consumption of unpaid music are hard to obtain. Moreover, even with data on volumes of paid and unpaid music consumption at the product level, it's difficult to identify the causal impact of stealing on buying. Works that are popular to buy are also popular to steal (Oberholzer-Gee and Strumpf 2007; Blackburn 2004), but the positive correlation between how much any work is stolen and how much it is purchased clearly does not prove that stealing music causes additional buying of that music.

These difficulties led a number of researchers to survey-based individual data, which seeks to elicit individuals' responses on their volumes of purchases and unpaid consumption of musical works. Studies of this nature, some of which examine within-individual variation across vintages, tend to find that stealing does depress buying (Zentner 2006; Rob and Waldfogel 2006; Waldfogel 2010). As the evidence has developed, most scholars in this area now agree that the unpaid consumption made possible by digitization is responsible for the lion's share of the revenue reduction in the music industry (for example, Liebowitz 2016).

The recorded music industry, which faced the toughest test because it was the first of the media industries to face digitization, did not develop an attractive legal method of digital distribution until Apple created the iTunes Music Store, four years after Napster (Knopper 2009). While this response may have been reasonably prompt by the standards of some industries, the four years between Napster and iTunes allowed consumers to grow accustomed to obtaining music without payment.

While the threat to revenue arising from the prospect of digital piracy exists for audio, video, and text—and hence for music, movies, television, and books—neither

movies, books, nor television have experienced a revenue decline resembling the collapse of recorded music revenue. The motion picture industry may have been protected from an explosion of piracy by the fact that video files are larger and more cumbersome to download than audio files. But in addition, adjacent media industries may have learned from this earlier experience that adapting to the new digital formats and searching for alternative revenue sources was more effective than trying to block them. YouTube appeared in February 2005, and people began uploading copyright-protected *Daily Show* clips without permission. While Viacom, the parent of Comedy Central, sued YouTube's parent Google for $1 billion, most television broadcasters embraced digital technology. By fall 2005, less than a year after YouTube's appearance, a few networks were posting episodes online at their own websites. By fall 2006, virtually all shows were posted online, free to consumers, for a few weeks after airing (Waldfogel 2009). Consumption of digital books requires complementary hardware, and the development of Amazon's Kindle reader was accompanied by the widespread availability of reasonably priced books at Amazon. Similarly, the growth in streaming video platforms such as Netflix and Amazon Instant has been accompanied by complementary technology for delivering video content to large television screens opposite couches, and not simply computers and phones. We will return later to the question of how digitization may offer additional revenue flows for media industries.

Digitization, Round 2: Falling Costs and Growth in the Number of New Products

While only some media industries—recorded music as well as newspapers—have faced the bad news of reduced revenue, all media industries experienced good technological news in the form of cost reduction. That is, digitization has brought substantial reductions in the costs of production, distribution, and promotion of new products in music, books, movies, and television. As a result, the gatekeeping role of media companies has been democratized.

The traditional model for bringing recorded music products to market involved several steps. First, a record label needed to identify a promising artist and sign the artist to the label. Second, the label spent substantial sums of money producing a recording, using expensive recording equipment and skilled workers. Next, the label produced a music video for television and secured the airplay of songs on the radio. Finally, the label had the album physically produced and shipped to stores. Because demand for popular music is often ephemeral, it was important for the product to be readily available during the few weeks it might be in high demand. The International Federation of the Phonographic Industries (2010) estimates that this mode of bringing music to market costs $1 million for an album from a new artist. And most releases were commercially unsuccessful (Vogel 2007; Caves 2000).

Digitization has offered low-cost alternatives to many of the steps in bringing products to market. Production is now far less expensive. An artist can create a

passable recording with an inexpensive microphone and the software on a computer or even a smart-phone. Distribution can be entirely digital. For about $10, an artist can make a song available on iTunes (Waldfogel 2015). Promotion remains a challenge, but many outlets review new music, including old-economy magazines such as *Rolling Stone* as well as born-online outlets such as Pitchfork. These reviews are available online, free of charge, and collected at sites such as Metacritic. Metacritic contains reviews of about 1,000 albums a year, whereas traditional terrestrial radio gives broad exposure to about 300 artists per year (Waldfogel 2015). Moreover, artists have opportunities to promote their work outside terrestrial radio, using YouTube, or online radio services such as Spotify and Pandora. There is some question about whether these outlets serve as demand-stimulating advertising or demand-depressing alternatives to the purchase of permanent download (as I discuss below in the context of revenue opportunities from bundling). However, rights-holders do get some payment for the use of their music via these outlets (for a sampling of the arguments on these issues, see David Lowery's blog at https://thetrichordist.com).

Gatekeeping roles have been transformed in other media industries as well. In the book publishing industry before digitization, and even as late as 2006, an author hoping for commercial success needed an agent who could help convince an editor at a major publishing house to publish the book (the discussion here draws on Waldfogel and Reimers 2015). Most manuscripts were rejected by agents. Publishers also screened the works, did some editing, and then had books printed and shipped to stores. Some books produced in this manner achieved sales success, although most did not. Publishers also provided information to taste-making critics, who might then provide reviews. With the development and widespread adoption of the e-book in 2007, these arrangements shifted. Since 2007, it has been possible for authors to create manuscripts, upload them to Amazon's Kindle Direct Publishing platform (or one of a number of others, such as Lulu) and then achieve multinational distribution without gatekeeping agents, editors, or publishers. Authors have availed themselves of these opportunities for self-publishing in substantial numbers.

As with other media industries, book promotion by individual or small-scale producers remains challenging. But digitization has sharply enriched the information environment in books. While traditional publications (newspapers and magazines) collectively provided roughly 50,000 reviews per year—with many books reviewed by multiple outlets and with the largest outlets reviewing about 8,000 titles—the number of books reviewed and rated on digital platforms is far larger. The two largest repositories of customer rating information about books are the user-generated review site Goodreads (with 10 million user reviews of 700,000 titles as of 2014) and Amazon's site. In fact, Amazon purchased Goodreads in 2014, making Amazon the owner of an enormous and difficult-to-imitate trove of review information.

Digitization has had similarly disruptive effects on the movie industry (the discussion here draws on Waldfogel 2016). The vast majority of the revenue for the movie industry has traditionally come from the major Hollywood studios, which

make up the membership of the Motion Picture Association of America (MPAA). The MPAA members have released between 150 and 250 movies per year into theaters over the past quarter-century. These movies have been quite expensive to make, averaging $106 million per film in 2007, the last year the MPAA released statistics.

The traditional model for movie distribution was to invest a large amount in a film thought to have promise, using well-known actors paid as much as $20 million or more for a single film as well as expensive cameras and equipment, and then to distribute the product in physical form to movie theaters around the world. But digitization has dramatically altered the parameters of both cost and distribution. Since about 2005, the cost of making a distribution-quality movie has fallen drastically. Digital SLR (single-lens reflex) cameras, using interchangeable lenses and capable of shooting high-definition video, have become available for a few thousand dollars, which is roughly 1 percent of the price of pre-existing distribution-quality film cameras. The reduction in the cost of this input is not material to a $100 million production budget, but it does enable filmmakers who lack MPAA-level financing to create professional-looking movies. On the distribution side, the number of physical theaters put a sharp constraint on the number of films that could effectively be distributed to consumers in the past. However, digitization has brought many new distribution channels, including video-on-demand through cable television operators, as well as pure online distribution through subscription platforms such as Netflix, Hulu, or Amazon Prime or a la carte platforms such as Amazon Instant.

In 2012, 550 films were distributed through US theaters: about 200 of these were MPAA major-studio movies, while the other 350 were small-scale releases of mostly independent movies. In many cases, these films were released briefly in just a few theaters to get some reviews, then, later, distributed through other channels. As of 2013, the number of 2010 releases available streaming on Netflix was 1,058, and the number on the Amazon Instant service was 1,230, or roughly twice the number of 2010 movies that had been available in theaters. The bottom line is that the barriers to entry into creation have fallen and the distribution bottleneck has been relaxed, making it possible for a large number of new movies to make their way to consumers.

Conditions for television programming are similar (the discussion here draws on Waldfogel 2017). As in the movie context, the physical costs of production (like cameras and recording media) have fallen substantially while channel capacity has grown enormously. Between the dawn of television and about 1990, the three national networks could accommodate about 25 new television series per year. With the growth of basic cable channels in the 1990s, the distribution capacity for new shows expanded. With the diffusion of digital cable systems, channel capacity grew to roughly 150 channels (Waldfogel 2017). Finally, the growth of the high-speed Internet has enabled asynchronous online distribution of new and old programs online via platforms such as Netflix, Hulu, and Amazon. Channel capacity is now effectively unlimited.

Even the Losers Get Lucky Sometimes: "Nobody Knows Anything" and the Welfare Benefit of New Products

Reductions in the cost of bringing products to market or making them available to consumers will increase the number of options facing consumers. One common metaphor used to describe the increased choices that consumers face since digitization is "the long tail" (for an academic treatment, see Brynjolffson, Hu, and Smith 2003; for a popular account, see Anderson 2006). The idea is well-illustrated by a comparison between the welfare consumers derived from, say, the 50,000 titles available in their local book stores compared with the 1,000,000 titles available to them from a retailer like Amazon that effectively has infinite shelf space. While each of the additional 950,000 titles has low demand, the sum of the incremental welfare delivered by many small things may be large. Brynjolffson, Hu, and Smith (2003) estimate that US consumers derived $1 billion in annual consumer surplus from the wide online selection. They derive this conclusion from the substantial share of book sales accruing to "long tail" books available at Amazon but presumed to be unavailable at a typical local store.

While the "infinite shelf space" perspective on the effect of digitization on consumer welfare is instructive, I wish to emphasize a separate effect of digitization on consumer welfare, drawing on Aguiar and Waldfogel (forthcoming). The reduction in the cost of bringing new products to market outlined above not only makes it possible for retailers to *carry* additional products; it also allows creators to *make* more products. It turns out that the unpredictability of appeal of new products has a large impact on the welfare benefit of new products.

To understand this point, first consider a world in which the quality of new products is entirely predictable at the time of initial investment. Suppose that gatekeepers (like record labels, movie executives, and book publishers) hear pitches and form estimates of expected revenue y' from prospective projects. If the estimate of expected revenue exceeds the cost threshold T, then the project goes forward. Otherwise the project does not. Initially, suppose that gatekeepers can forecast revenue without error. Then all projects with expected revenue $y' > T$ are released. When digitization reduces the cost of launching products from T to T', then more products get produced and released. Consumers get access to new products, but all of these new products have lower expected *and* realized appeal than the least marketable product previously made available. In this circumstance, with perfectly predictable product quality, the additional product releases made possible by cost-reducing digitization increase consumer welfare in the same way that adding shelf space raises consumer well-being.

However, the unpredictability of outcomes raises the possibility that a cost reduction that stimulates new products will deliver a much bigger welfare benefit. To see this, suppose that gatekeepers are unable to forecast market appeal with complete accuracy. Then their guess is the true value plus an error term, or $y' = y + \varepsilon$. Because of unpredictability, some projects have realized marketability above the initial cost threshold T; others below. When digitization reduces the

entry threshold from T to T', additional projects with modest expected revenues—expected marketability between T and T'—are brought to market. But because of unpredictability, high realized quality appears throughout the distribution of projects, many of which would previously have been rejected by gatekeepers. Some of the products that had been expected to "lose" instead turn out to be winners, in the form of best-selling products.

Unpredictability is a generic feature of creative products, as Caves (2000) and Vogel (2007) emphasize, with evidence that roughly 5–10 percent of new creative products achieve success in the sense of generating revenue in excess of their costs. As screenwriter William Goldman (1984) said in his *Adventures in the Screen Trade:* "NOBODY KNOWS ANYTHING." Hence, we should expect digitization to have a big effect on the welfare benefit that consumers derive from growth in the number of new products.

In short, cost reduction allows creators to take more draws from a lottery of possible winners. Given unpredictability, some proportion of the additional draws will deliver some additional high-quality products, which, in turn, could raise the quality of the choice set facing consumers. If correct, this characterization has three empirical implications: 1) the number of new products will rise; 2) the service flow that consumers derive from the new vintages will rise; and 3) many new products expected to have been losers, in the sense that they would not have made it past the traditional gatekeepers, will make up a growing share of the actual winners in these markets. The next sections explore these three implications.

Evidence of a Digital Renaissance

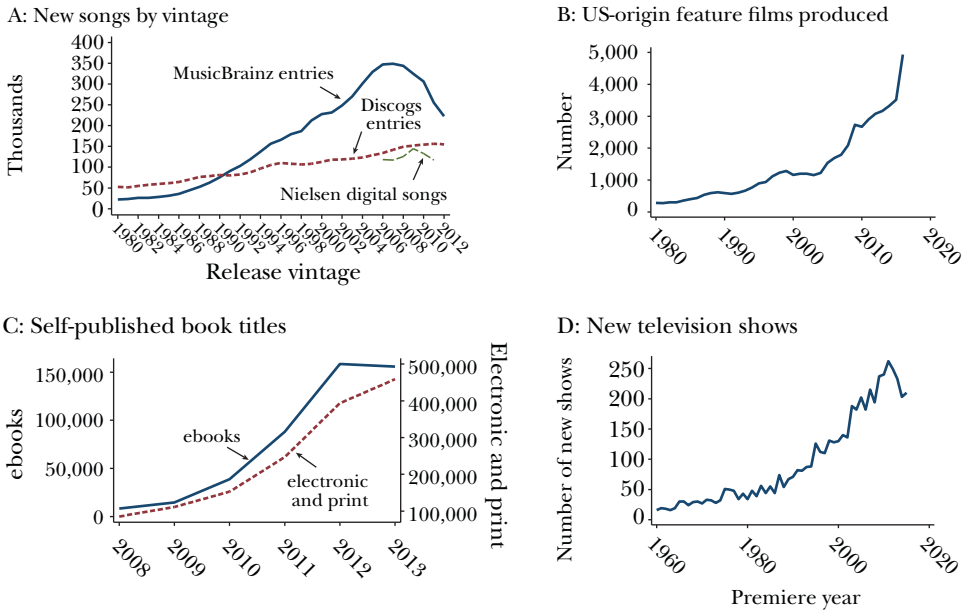
The Number of New Products

While obtaining data on the number of new products can be challenging, it is clear that the number of products brought to market has grown sharply in each of the media product contexts. Figure 2 presents some measures of the numbers of new products brought to market in the US in each of four industries.

Figure 2A shows that the estimate of the number of new recorded music works released differ across sources, but all agree that the number has risen since about 2000. According to entries in the Musicbrainz database, the number of new songs added annually rose steadily from about 50,000 in 1988 to almost 350,000 in 2007. Other reports put the growth in the number of new recorded musical works at a tripling between 2000 and 2010 (for details, see Aguiar and Waldfogel 2016, forthcoming).

The growth in motion picture production is similarly large (Figure 2B). Based on production data in the Internet Movie Database (IMDb), the number of new motion pictures produced in the United States rose from about 500 features in 1990 to 1,200 in 2000, and by 2010 had risen to nearly 3,000. Growth in US-origin documentaries is even larger, and the patterns for other countries are similar (Waldfogel

Figure 2
Number of Products Brought to Market in Four Media Industries



Source: The data on new songs comes from MusicBrainz, Discogs, and Neilson; on self-published books, from Bowker; on feature films, from IMDb; and on television shows, from epguides.com.

2016). Not all of the new movies are marketed to consumers, but even the number distributed to consumers through some familiar distribution channel—theaters, Amazon, Netflix, iTunes—has at least doubled between 2000 and 2010.

The growth in new books has been even larger (see Figure 2C). Much of the growth in new books has come from self-published works, and their growth has been substantial, from about 85,000 new titles in 2008 to almost 400,000 new titles in 2012 (reading off the right axis) (Waldfogel and Reimers 2015).

Finally, Figure 2D shows the number of new US television series grew from about 25 to 50 over 1960–1980, rose to 100 by 2000, and has since topped 250 (according to epguides.com). Production-based estimates derived from IMDb show the same time pattern, but roughly ten times the level (Waldfogel 2017).

It is clear that the numbers of new musical works, movies, books, and television shows created and made available to consumers have risen sharply since digitization.

The Service Flow Delivered by New Products

Of course, the large number of new products does not guarantee or even imply any substantial growth in the service flow delivered by the new vintages. Some cultural critics have decried an onslaught of amateurish cultural products (Lemann 2006). For example, one of the fruits of digitization is the growth in the number of YouTube videos depicting cats on Roombas—a brand of robot vacuum cleaner that

navigates autonomously around a room.[1] Still, it is possible that the large recent crops of cultural products include valuable entries.

Determining the welfare benefit that consumers derive from the new works requires some assessment of the quality of the new works. "Quality" is a loaded word for creative products, so it is helpful to be more specific. I have measured the service flow of new cultural products based on consumers' buying decisions, and where that is not possible, I measure quality based on the judgment of critics. I am particularly interested in trends in quality since the introduction of Napster in 1999 as indicated by a vertical line in the panels of Figure 3.

A first critic-based approach employs intertemporally comparable assessments. One form of this approach is multiyear "best of" lists, such as the *Rolling Stone* list of the best 500 albums of all time. In Waldfogel (2012), I create indices from a few dozen underlying music best-of lists, which I combine via a regression to create an index of the number of high-quality releases from each vintage, 1960–2008. I report the resulting index of critically acclaimed music in Figure 3A. The index rises from 1960 to 1970, and then falls. The index also declines from a local peak in the mid-1990s toward 1999. In stark contrast with the time pattern of revenue, which collapsed after Napster, the index of acclaimed music is flat. At least by this measure, the decline in revenue has not undermined the creation of new music that critics find appealing. We return to the evolution of music quality shortly, with usage-based evidence.

A related type of critic-based information comes from information intermediaries that translate movie or music reviews into scores on a 0–100 scale, such as Rotten Tomatoes or Metacritic. In some recent research (Waldfogel 2016, 2017), I create indices showing the number of movies and television shows, respectively, with ratings above some fixed threshold of critical acclaim. For example, 1993 saw the release of fewer than 40 films receiving scores of 87 or higher on Rotten Tomatoes. Since about 2000, the number has grown substantially, averaging about 80 and reaching 100 in 2012. Hence, the quality of movies produced has been rising in the eyes of critics.

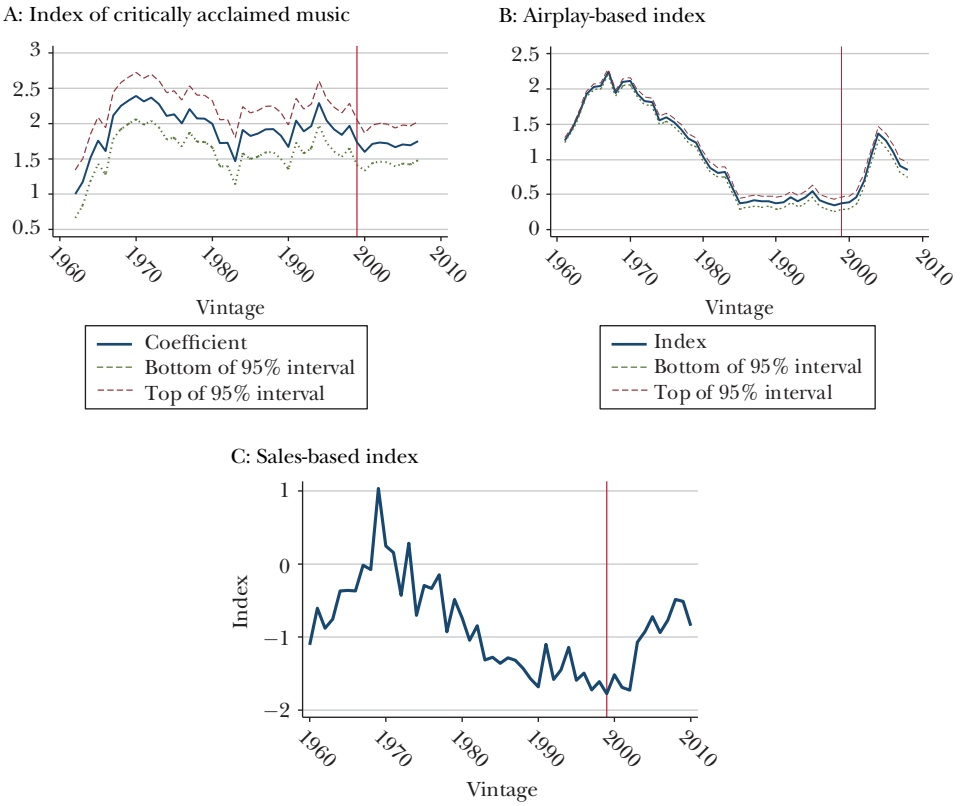
For the evolution of viewers' assessments of television shows released over the years, I employ user ratings of shows. IMDb users can rate television shows on a 10-point scale, and the IMDb database includes a comprehensive listing of the series premiering each year. An analysis of 13,439 US series premiering between 1970 and 2015 shows rising variance in user-assessed series quality over time, which is highly consistent with the unpredictability hypothesis entertained above. The quality of the shows that turn out (after being produced) to be most appealing to viewers rises. That is, the ratings of the top 25 premiering shows of each year, according to the ratings left by IMDb users, rise sharply over time. One sees a similar time pattern in professional critics' Metacritic ratings of the top 25 shows of the seasons from 2000 to 2015 (Waldfogel 2017). Indeed, the observation that we are currently living

[1] The "Roomba Cats: Compilation" at YouTube (https://www.youtube.com/watch?v=mk4XB2wZqF4) has nearly 500,000 views. I am grateful to Brett Danaher for this colorful reference.

Figure 3

Indices of the Quality of New Music

(the vertical line indicates the creation of Napster)



Source: Figure 3A above corresponds to Waldfogel (2012, figure 3); Figure 3B corresponds to Waldfogel (2012, figure 8); Figure 3C corresponds to Waldfogel (2012, figure 10).

through a television Golden Age is not novel; many observers speak of "peak TV" (for example, Poniewozik 2015).

The second broad approach to assessing the evolution of quality relies on consumers' choices. At any point in time, consumers can choose either new or old products. Because consumers are generally less enthusiastic about older products than new ones—a phenomenon akin to depreciation—new products tend to be used more at any point in time. This observation gives rise to a way of assessing quality: after accounting for age, are some vintages of, say, music, used more intensively than others? Implementing this approach requires data on consumption of products by calendar time and vintage for multiple calendar years. That is, one needs data on the share of music consumption in 2010 that was music originally released in 2010, 2009, and so on. If one can observe the same thing for calendar 2009 as well as earlier years, then it's possible to measure directly

whether some vintages are used more or less than others, after accounting for depreciation.

I have implemented this approach using several different datasets, including one based on airplay of US music, 2004–2008, and another one using Gold and Platinum record certifications certifying 0.5 million or sales multiplies of one million between 1970 and the early 2000s (Waldfogel 2012). Figure 3B and C report the resulting indices. First, as with the independently derived index based on critical acclaim, both of the usage-based indices rise from about 1960 to 1970, then decline. Both have minor fluctuations during the 1990s. However, both rise rather sharply after 1999. These indices indicate that the vintages of music released since Napster are more used, conditional on their age, than the previous vintages. To say this another way, the service flow of the new vintages has risen relative to the vintages of the 1990s.

Some of the methods for measuring the evolution of quality of content are subject to a potential bias toward particular vintages. For example, the views of critical judgments from a particular year might over-value recent work. But some of the methods and data I employ to assess the evolution of quality are immune to this concern: for example, an approach that infers the evolution of music quality from the sales certifications—for albums selling over half a million, and multiples of one million copies—occurring over four decades. While the sales data from, say, a particular decade would reflect the tastes of that decade's buyers, potentially over-weighting whatever music was popular in that decade, having four decades of sales-based certifications means that the inferences about vintage quality are based on buying behavior over a long period.

The Growing Role of Those Who Expected to Lose, But Ended up Winning

Growth in the experienced quality of new products does not by itself demonstrate that digitization delivered these benefits. After all, quality improvements might have occurred even without the new products. We can explore whether digitization is responsible for the quality of new products by documenting the role of new products among the products that consumers find most appealing. The test for whether digitization is responsible is whether the products with modest expected appeal—or those that would have been expected to be "losers" and thus unlikely to be produced at all at an earlier time—account for substantial and growing shares of the actual winners.

The first task is to identify the output that would have been an expected loser. In the terminology introduced earlier, expected losers are products with expected revenue below the old green-lighting threshold and above the new, lower one (in the terms used earlier, between T and T'), which would not have come to market but for digitization. They are the rejected manuscripts, demo tapes, and story pitches, the would-be products that gatekeepers would have scotched if costs had remained high. A short digression into the markets for music, books, movies, and television provides reasonable guidance on how to identify these expected losers.

The entities traditionally bringing new music to market are divided into two main groups: "major" record labels and "independents." The majors are owned by major media conglomerates, firms like Warner, Sony, and Universal. While major labels have always issued only a small fraction of the total releases, major label products have accounted for the vast majority of sales—roughly 90 percent as late as 2000. Artists who could obtain major label deals—with their substantial advances, access to radio airplay, and broad distribution—would generally jump at these opportunities. Some artists unable to obtain major label deals could nevertheless obtain independent label deals. In the past 16 years, as Handke (2012) and Oberholzer-Gee and Strumpf (2010) document, there has been substantial growth in independent labels and, more importantly, in releases by independent labels. Empirically, we can treat the releases on independent record labels as the products that would have been expected to be commercial losers by the major labels. In Waldfogel (2012, 2015) and Aguiar and Waldfogel (2016, forthcoming), I ask whether these releases account for a growing share of bestselling works.

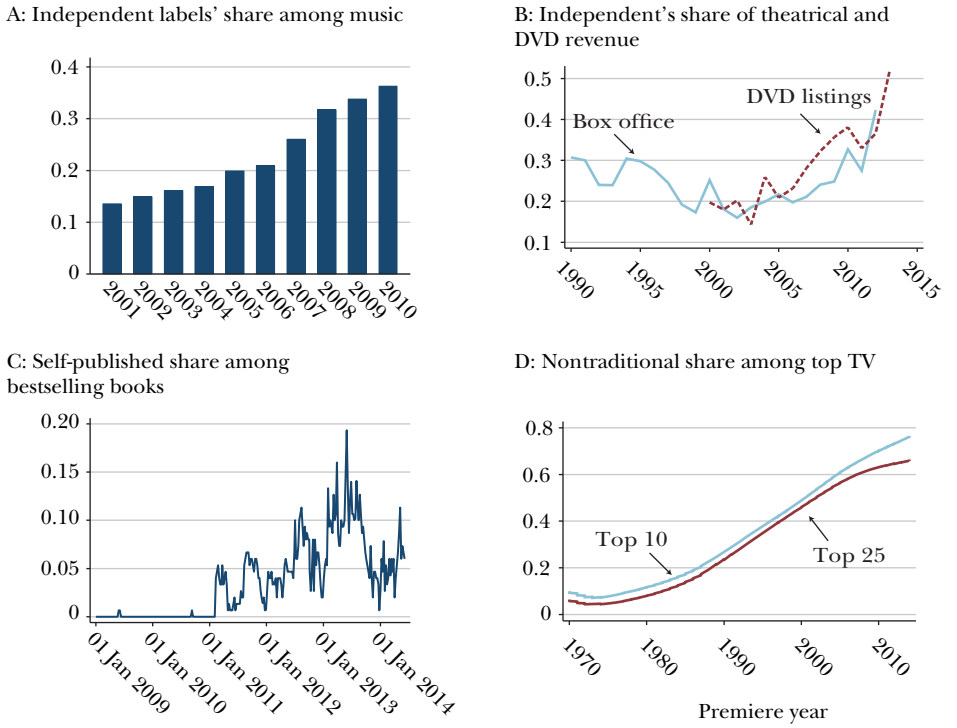
We can make a similar distinction in the market for books. Traditionally, the books rejected by mainstream publishers either remained in desk drawers, or perhaps sometimes were self-published by "vanity presses." Today, as detailed above, many books whose authors cannot secure deals with publishers self-publish their books through Amazon and other platforms. The question, then, is whether such self-published works account for a growing share of best-selling works. I undertake this calculation in Waldfogel and Reimers (2015).

The movie industry, like the recording industry, is divided into major and independent studios. With movies, we can ask what share of box office revenue, or share of works distributed, hail from independent as opposed to major studio sources (as in Waldfogel 2016).

In television, the question is whether shows not originating with the traditionally powerful gatekeepers—the legacy broadcast networks and HBO—account for a growing share of the most popular shows. Thus, I treat shows that premiered outside these traditional channels as expected losers, a theme I explore in Waldfogel (2017).

Figure 4 shows various measures of the share of successful products accounted for by products that apparently were perceived as having only modest promise, because they were not produced by the traditional media gatekeepers. In the recorded music industry (panel A), the share of top-selling albums released by independent labels grew from about 12 to 35 percent between 2000 and 2010 (Waldfogel 2015). The motion picture industry has seen a similar transformation, as shown in panel B: between 2000 and 2012, the share of box office and DVD revenue accounted for by independent movies grew from 20 to about 40 percent (Waldfogel 2016). In the book industry (panel C), between the appearance of the Kindle in 2007 and 2014, the share of best-selling books that originated as self-published works grew from zero to over 10 percent. In the romance category, the share topped 40 percent by 2014 (Waldfogel and Reimers 2015). In television (panel D), the share of the shows originating outside of the traditional sources (major broadcast networks

Figure 4
The Share of Expected Losers among the Actual Winners



Sources: Panel A: Waldfogel (2015); panel B: Waldfogel (2016a); panel C: Waldfogel and Reimers (2015); panel D: Waldfogel (2017).

or HBO) among those shows rated most highly, grew from under 10 percent in 1970 to about 75 percent by 2010.

In short, in all four industries, the expected losers have rising and now substantial shares. These patterns are consistent with the belief that many of the new media products that are valued and purchased by consumers would have been unavailable to consumers absent digitization.

Quantifying the Welfare Benefit of Digitization

How large are digitization's benefits to consumers? The conventional "long tail" view of digitization of the media industries focuses on retailing. Rather than the small number of book, movie, or music titles available to consumers at a local bricks and mortar store, online retailing allows consumers access to essentially all extant products. This has clearly delivered large benefits to consumers, particularly those facing limited offline choices; and as noted earlier, Brynjolfsson, Hu, and Smith (2003) estimate that US consumers derive $1 billion in annual additional consumer surplus from access to the full online, rather than the limited offline, book choice set.

But there is another way to think about digitization's benefit to consumers. By reducing the cost of bringing new products to market, digitization has enabled entry of large numbers of new products. Because product quality is unpredictable, some of these new products have turned out to be quite good. Aguiar and Waldfogel (forthcoming) quantify the size of the welfare benefit arising from a tripling in the volume of new music releases associated with this "random long tail" in recorded music, estimating it to be roughly 20 times larger than simply the benefit of giving consumers access to a longer tail of low-value products.

The intuition behind the finding works like this. If the quality of new media products were completely predictable at the time of green-lighting decisions, then a cost reduction that, say, tripled the number of new products would benefit consumers. But the benefits would be small, in the sense that all of the new products would be less appealing than the least-appealing product that was previously viable. The benefit would be equivalent to the benefit of getting access to, say, a large number of low-demand products too unpopular to be stocked at local stores.

Suppose instead—and in line with how creative products work—that producers cannot perfectly predict the appeal of creative products at the time of investment. Then a cost reduction that triples the number of new products brought to market will bring forth a range of product qualities, some good, some bad. If we order products according to their expected appeal before production, then some of the "ex ante losers" turn out to be "ex post winners"—that is, products that would not have made the gatekeepers' cut to be produced before digitization turn out to be highly desired by consumers. Indeed, given this unpredictability of appeal for media products, we estimate that the bottom two-thirds of new music products according to expected appeal accounts for about 20 times more sales than the bottom two-thirds of products according to actually realized appeal for consumers. Hence, we conclude that the additional products made possible by digitization, which randomly includes a large number of expectation-beating entries in quality, is large compared to the conventional long tail that just offers more choices, and that as a result, digitization has had a large effect on the welfare of consumers.

Two other aspects of the welfare effects of digitization bear discussion. First, digitization has a separate effect operating through reduced costs of maintaining inventories. The unpredictability of product appeal made retailing expensive prior to digitization. The music-, movie-, and bookstores needed to stock products for which demand might or—more often—might not materialize. With digital distribution, consumers have access to large selections without requiring a costly-to-maintain retail sector. Further distribution is global: consumers everywhere get access to most of the same products.

Second, despite the benefits of digitization operating through the creation of more and sometimes better products, the additional products bring with them an additional cost of discovering which of the new products are worthy of attention. This cost is exacerbated by the fact that cultural products are experience

goods. More research is needed in this area, but some comments are in order. First, it is not clear, even with many more products, that social product discovery costs have risen. In the case of music, product discovery traditionally operated through songs being aired on radio; an entire radio station audience (of, say, 100,000 listeners) needed to sample a song each time the station aired it. After airing a song, say, 10 times, a station might learn whether the song would be appealing to its listeners. With those hypothetical numbers, the cost of testing that song would be 1,000,000 listens. In the digital environment, by contrast, song sampling can occur one listener at a time; and it's possible that a song could be "discovered" to be appealing (based on listening occurring at streaming sites) based on far fewer listens. Even with far more products coming to market, the social cost of product discovery may have fallen. Having said this, it is also possible that product discovery in media markets is subject to informational cascades (Bikhchandani, Hirshleifer, and Welch 1998), in which products that get off to a strong start develop a "cascade" of positive feedback that carries them forward—and vice versa for products that get off to a slower start. Recent empirical work documents patterns of observational learning in music markets (Newberry 2016).

Digitization and Revenue Opportunities

In addition to reducing the cost of bringing new products to market, digitization has also created some new revenue opportunities. First, by turning creative works into zero marginal cost media products, digitization has facilitated the use of bundled selling arrangements. Second, zero marginal cost distribution has also, in some media industries, created revenue opportunities from the sale of complements to the digital products, such as live musical performances.

The Promise of Bundled Sales of Zero Marginal Cost Products

One effect of digitization when applied to media products is that the marginal cost of serving another consumer falls essentially to zero. When this change is combined with the ability of people to access media products on a wide range of home and portable devices, digitization enables a new range of sophisticated sales and pricing strategies, which at least in theory could bring revenue benefits to sellers. Here, we will particularly focus on the possibilities of bundled sales strategies, which have particular advantages when 1) products have zero marginal costs, and 2) consumers' valuations of products are not (perfectly) positively correlated. The media products discussed here have these characteristics: all can be digitally transmitted at essentially zero marginal cost, and different consumers attach high value to different products.

In the music industry, sales of downloaded music seemed to take off in 2003 when Apple's iTunes Music store launched, charging a uniform $0.99 price (for most content in the United States). Starting in 2009, the store moved to three

tiers at $0.69, $0.99, and $1.29 (Apple 2009). Both theory (such as Bakos and Brynjolffson 1999) and empirical research on music (Shiller and Waldfogel 2011) had suggested the promise of more sophisticated pricing strategies for digital products, and there has been a move recently to bundled content through music-streaming services. For example, Spotify, which offers bundled access to a wide variety of music, has grown from 0.5 million paid subscribers in 2010 to 50 million paid subscribers in March 2017 (see https://www.statista.com/statistics/244995/number-of-paying-spotify-subscribers/).

Video—in the form of both movies and television—is now commonly distributed to the home market in bundles. Netflix is perhaps the most prominent example, digitally delivering 4,210 movies and 798 television series in the United States for as little as $7.99 per month (see https://www.justwatch.com/us/provider/netflix?content_type=show). While Netflix has a reasonably large library of already-existing movies and shows, it includes relatively few recent blockbusters. Netflix has begun to produce original programming, such as *House of Cards* and *Orange is the New Black*. Platforms such as Hulu and Amazon's Prime service also offer bundles of programming for flat fees.

Providers are experimenting with book bundles as well, including Amazon's Kindle Unlimited service, as well as Scribd and other services.

Whether a combination of bundling and streaming stimulates or depresses other recorded music revenue is an important topic for the content industries. Concerned that their payments from streaming on Spotify would not compensate them for depressed album sales, the biggest-selling artists of the past few years—Taylor Swift and Adele—withheld their new albums from Spotify. While Spotify includes a massive library, these prominent defections raise the possibility that streaming services often offering bundles will exclude new, high-value content. This pattern is common in media industries: for example, movies first show in first-run theaters and then in second-run theaters, and then become available online. Books first appear in hardcover editions, and later in paperback. It is possible that streaming services will be part of a similar pattern of inter-temporal price discrimination, as a mode of distribution and source of revenue after the willingness to pay of high-valuation consumers has been harvested. Various recent papers have documented that streaming does displace sales: for example, Wlömert and Papies (2016) use individual-level survey data, and Aguiar and Waldfogel (2015) use aggregate data. Overall, impacts of streaming on rights-holder revenue depend on per-stream payments, which are the subject of ongoing industry discussions.

Recorded music and live music are complements in the view of many fans. The ability to distribute digital copies of recorded music at zero marginal cost raises the possibility of using recorded music as advertisements for live performances. Mortimer, Nosko, and Sorenson (2012) document that digitization stimulates concert ticket sales, at least for relatively obscure artists. Connolly and Krueger (2006) document that concert ticket prices have risen since digitization.

Conclusion

Digitization arrived to many industries as revenue-reducing bad news. Yet the main effect of digitization—even in the industries such as music, which has seen a catastrophic decline in revenue—has been to reduce the cost of bringing new works to market. While it may be true that some industry participants face challenges from digitization, such as traditional major labels, studios, and publishing houses, it also seems clearly true that consumers are now awash in products that they find desirable. To put it succinctly, digitization has ushered in a golden age of music, movies, books, and television programming.

While the early views of digitization's effect on consumers in media markets—adding access to the long tail of existing products—is correct, and these welfare gains are substantial, the effects of digitization on production are even more substantial. The lessons about the impact of digitization on the benefits from new products may have application outside of media markets. Whenever the appeal of new products is unpredictable, reduction in the cost of bringing new products to market holds out the promise that new products will deliver substantial improvements. New product quality is understood to be unpredictable in many industries, so these ideas may apply more broadly.

Much of the public policy response to digitization has concerned methods for curbing piracy, ostensibly because piracy threatens continued investment in content. While it is clearly correct in principle that threats to revenue, all else constant, stand as threats to continued content creation, it is also true that threats to revenue in media industries have been accompanied by reductions in cost. Assessments of whether copyright is fulfilling its function require more than documenting that revenue has fallen. Instead, assessments of copyright should be based on the evidence of new content creation. There can be good reasons to enforce rules against piracy; after all, stealing is illegal. But a crisis in realized creative output since 1999 should not be among the reasons cited for strengthening effective intellectual property protection for media products.

The topic of digitization and its effects on content industries is a fertile area for further research. Among the particularly important questions are product discovery, which refers to the ways in which consumers sift through the large number of new products to find those that they find appealing, as well as the effects of global distribution (for example, via Netflix, Amazon, Spotify, and iTunes) on the supply of new products and the possible convergence of consumption across places.

References

Aguiar, Luis, and Joel Waldfogel. 2015. "Streaming Reaches Flood Stage: Does Spotify Stimulate or Depress Music Sales?" NBER Working Paper 21653.

Aguiar, Luis, and Joel Waldfogel. 2016. "Even the Losers Get Lucky Sometimes: New Products and the Evolution of Music Quality since Napster." *Information Economics and Policy* 34: 1–15.

Aguiar, Luis, and Joel Waldfogel. Forthcoming. "Quality Predictability and the Welfare Benefits from New Products: Evidence from the Digitization of Recorded Music." *Journal of Political Economy.*

Anderson, Chris. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More.* Hyperion Books.

Apple. 2009. "Changes Coming to the iTunes Store." https://www.apple.com/pr/library/2009/01/06Changes-Coming-to-the-iTunes-Store.html.

Bakos, Yannis, and Erik Brynjolfsson. 1999. "Bundling Information Goods: Pricing, Profits, and Efficiency." *Management Science* 45(12): 1613–30.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1998. "Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades." *Journal of Economic Perspectives* 12(3): 151–70.

Blackburn, David. 2004. "On-line Piracy and Recorded Music Sales." Unpublished manuscript.

Brynjolfsson, Erik, Yu (Jeffrey) Hu, and Michael D. Smith. 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers." *Management Science* 49(11): 1580–96.

Caves, Richard E. 2000. *Creative Industries: Contracts between Art and Commerce.* Cambridge: Harvard University Press.

Connolly, Marie, and Alan Krueger. 2006. "Rockonomics: The Economics of Popular Music." In *Handbook of the Economics of Art and Culture,* Vol. 1, edited by Victor Ginsburgh and David Throsby, 667–719. Elsevier.

Goldman, William. 1984. *Adventures in the Screen Trade: A Personal View of Hollywood and Screenwriting.* Grand Central Publishing.

Handke, Christian. 2006. "Plain Destruction or Creative Destruction? Copyright Erosion and the Evolution of the Record Industry." *Review of Economic Research on Copyright Issues* 3(2): 29–51.

International Federation of the Phonographic Industry (IFPI). 2010. "Investing in Music: How Music Companies Discover, Develop, and Promote Talent." https://web.archive.org/web/20101126054054/http://ifpi.org/content/library/investing_in_music.pdf.

Knopper, Steve. 2009. *Appetite for Destruction: The Spectacular Crash of the Record Industry in the Digital Age.* Free Press.

Lemann, Nicholas. 2006. "Amateur Hour: Journalism without Journalists." *New Yorker,* August 7. http://www.newyorker.com/magazine/2006/08/07/amateur-hour-4.

Liebowitz, Stan J. 2004. "The Elusive Symbiosis: The Impact of Radio on the Record Industry." *Review of Economic Research on Copyright Issues* 1(1): 93–118.

Liebowitz, Stan J. 2016. "How Much of the Decline in Sound Recording Sales Is Due to File-Sharing?" *Journal of Cultural Economics* 40(1): 13–28.

Mortimer, Julie Holland, Chris Nosko, and Alan Sorensen. 2012. "Supply Responses to Digital Distribution: Recorded Music and Live Performances." *Information Economics and Policy* 24(1): 3–14.

Newberry, Peter W. 2016. "An Empirical Study of Observational Learning." *RAND Journal of Economics* 47(2): 394–432.

Oberholzer-Gee, Felix, and Koleman Strumpf. 2007. "The Effect of File Sharing on Record Sales: An Empirical Analysis." *Journal of Political Economy* 115(1): 1–42.

Oberholzer-Gee, Felix, and Koleman Strumpf. 2010. "File Sharing and Copyright." *Innovation Policy and the Economy* 10(1): 19–55.

Poniewozik, James. 2015. "Emmy Awards 2015: A Show for a 'Peak TV,' Blockbuster Era." *New York Times,* September 21. https://www.nytimes.com/2015/09/21/arts/television/emmys-2015-andy-samberg-review.html.

Rob, Rafael, and Joel Waldfogel. 2006. "Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students." *Journal of Law and Economics* 49(1): 29–62.

Shiller, Ben, and Joel Waldfogel. 2011. "Music for a Song: An Empirical Look at Uniform Pricing and Its Alternatives." *Journal of Industrial Economics* 59(4): 630–60.

Vogel, Harold L. 2007. *Entertainment Industry Economics: A Guide for Financial Analysis.* 7th ed. Cambridge University Press.

Waldfogel, Joel. 2009. "Lost on the Web: Does Web Distribution Stimulate or Depress Television Viewing?" *Information Economics and Policy* 21(2): 158–68.

Waldfogel, Joel. 2010. "Music File Sharing and

Sales Displacement in the iTunes Era." *Information Economics and Policy* 22(4): 306–14.

Waldfogel, Joel. 2012. "Copyright Protection, Technological Change, and the Quality of New Products: Evidence from Recorded Music since Napster." *Journal of Law and Economics* 55(4): 715–40.

Waldfogel, Joel. 2015. "Digitization and the Quality of New Media Products: The Case of Music." In *Economic Analysis of the Digital Economy*, edited by Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, 407–42. National Bureau of Economic Research.

Waldfogel, Joel. 2016. "Cinematic Explosion: New Products, Unpredictability, and Realized Quality in the Digital Era." *Journal of Industrial Economics* 64(4): 755–72.

Waldfogel, Joel. 2017. "The Random Long Tail and the Golden Age of Television." In *Innovation Policy and the Economy*, vol. 17, edited by Shane Greenstein, Josh Lerner, and Scott Stern, 1–25. National Bureau of Economic Research.

Waldfogel, Joel, and Imke Reimers. 2015. "Storming the Gatekeepers: Digital Disintermediation in the Market for Books." *Information Economics and Policy* 31: 47–58.

Wlömert, Nils, and Dominik Papies. 2016. "On-Demand Streaming Services and Music Industry Revenues: Insights from Spotify's Market Entry." *International Journal of Research Marketing* 33(2): 314–27.

Zentner, Alejandro. 2006. "Measuring the Effect of File Sharing on Music Purchases." *Journal of Law and Economics* 49(1): 63–90.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 215–230

Retrospectives
Friedrich Hayek and the Market Algorithm

Samuel Bowles, Alan Kirman, and Rajiv Sethi

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 contact Joseph Persky, Professor of Economics, University of Illinois, Chicago, at jpersky@uic.edu.

Introduction

Friedrich A. Hayek (1899–1992) is known for his vision of the market economy as an information processing system characterized by *spontaneous order*: the emergence of coherence through the independent actions of large numbers of individuals, each with limited and local knowledge, coordinated by prices that arise from decentralized processes of competition. Hayek is also known for his advocacy of a broad range of free market policies and, indeed, considered the substantially unregulated market system to be superior to competing alternatives precisely because it made the best use of dispersed knowledge:

■ *Samuel Bowles is Research Professor and Director of the Behavioral Sciences Program, Santa Fe Institute, Santa Fe, New Mexico. Alan Kirman is Professor Emeritus of Economics, Ecole des Hautes Etudes en Sciences Sociales, Paris, France. Rajiv Sethi is Professor of Economics, Barnard College, Columbia University, New York City, New York, and External Professor, Santa Fe Institute, Santa Fe, New Mexico. Their email addresses are samuel.bowles@gmail.com, alan.kirman@ehess.fr, and rs328@columbia.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.215

doi=10.1257/jep.31.3.215

> [The market is] a system of the utilization of knowledge which nobody can possess as a whole, which … leads people to aim at the needs of people whom they do not know, make use of facilities about which they have no direct information; all this condensed in abstract signals … [T]hat our whole modern wealth and production could arise only thanks to this mechanism is, I believe, the basis not only of my economics but also much of my political views (Hayek 1994, p. 69).

These political views included opposition not only to Soviet-style central planning, but also to monetary and fiscal demand management policies, collective bargaining, wage floors, and significant public expenditures. Such forms of interference with the market, in his view, would compromise its ability to deliver continued prosperity.[1] His hostility to Keynes and Keynesian policies, in particular, was deep and visceral.

Following conventional usage, we shall use the term *laissez faire* to represent this general stance and the associated suite of policy positions. Hayek himself rejected the term, which he associated with a tradition in social thought that considered human beings to be endowed with the "intellectual and moral attributes" necessary to "fashion civilization deliberately" (Hayek 1960, pp. 60-61). He firmly opposed the view that institutions were "deliberate contrivances," arguing instead that they emerged through trial and error across generations. Successful societies were those in which "man's more primitive and ferocious instincts" were "tamed and checked by institutions that he neither had designed nor could control." These institutions would then survive and spread through learning and imitation rather than deliberate design.

Hayek drew a sharp contrast between his approach and Walrasian general equilibrium theory, which itself had been used to make a case for *laissez faire* on the basis of the two fundamental theorems of welfare economics. These can be roughly stated as follows: a competitive price-taking market equilibrium will be Pareto-efficient, and any distributional concerns about the outcomes of such a market can be addressed through a redistribution of endowments. It was these theorems that Gérard Debreu (1984) presumably had in mind when he reportedly claimed that "the superiority of the liberal economy is incontestable and can be mathematically demonstrated." In contrast, Hayek did not consider the welfare theorems to be compelling arguments for his policy stance. As he put it, the "argument in favor of competition does not rest on the conditions that would exist if it were perfect" (1948, p. 104). Instead, his case for competitive markets rested on the idea that competition was a "procedure for discovering facts which, if the procedure did not exist, would remain unknown or at least would not be used" (Hayek 1968). In this view, the superiority of competition as a procedure for discovering and utilizing knowledge could be established only through a comparative evaluation of economic systems.

[1] However, Hayek did support a universal basic income (1979, p. 55), and was generally opposed to free banking (White 1999).

Our purpose in writing this paper is twofold:

First, we believe that Hayek's economic vision and critique of equilibrium theory not only remain relevant, but apply with greater force as information has become ever more central to economic activity and the complexity of the information aggregation process has become increasingly apparent. Advances in computational capacity and the growth of online transactions and communication have made the collection and rapid processing of big data feasible and profitable. Many markets now involve algorithmic price-setting and order placement alongside direct human action, raising interesting new questions about the processes by which information is absorbed and transmitted by prices.

Second, we wish to call into question Hayek's belief that his advocacy of free market policies follows as a matter of logic from his economic vision. The very usefulness of prices (and other economic variables) as informative messages—which is the centerpiece of Hayek's economics—creates incentives to extract information from signals in ways that can be destabilizing. Markets can promote prosperity but can also generate crises. We will argue, accordingly, that a Hayekian understanding of the economy as an information-processing system does not support the type of policy positions that he favored. Thus, we find considerable lasting value in Hayek's economic analysis while nonetheless questioning the connection of this analysis to his political philosophy.

It is worth noting that Hayek shared the 1974 Nobel Memorial Prize in Economics with Gunnar Myrdal "for their penetrating analysis of the interdependence of economic, social and institutional phenomena." These two economists were poles apart politically, one being a committed social democrat and the other a classical liberal. Yet, if the argument in this paper is sound, Hayek's economic vision ought to be of value to those with Myrdal's politics, just as Myrdal's analytical contributions remain of broad interest and relevance.

Hayek on Competition, Equilibrium, and Disequilibrium

Even prior to the publication of his celebrated 1945 paper "The Use of Knowledge in Society," Hayek had developed a highly sophisticated and pioneering understanding of intertemporal equilibrium and the conditions under which it could be achieved or sustained. In a 1937 paper in *Economica*, he defined equilibrium as a set of individual plans that could be executed without mutual interference. This allows for the possibility that individual beliefs depend upon local knowledge and differ, provided that these beliefs are not contradicted as plans unfold. This notion of equilibrium is thoroughly modern, dynamic, and unrestrictive—and quite distinct from a general equilibrium model in which prices are uniform and public. Its development by Hayek is a significant—though little recognized—accomplishment in its own right. Indeed, Hayek claimed later in life that it "seems to me in

retrospect the most original contribution I have made to the theory of economics (1994, p. 68).[2]

But Hayek was not particularly interested in the properties of equilibrium itself, and saw the strength of the market economy as arising from the learning and diffusion of new information that it accomplishes in *disequilibrium*. Unforeseen (and often unforeseeable) changes in economic fundamentals that are initially recognized by only a small number of individuals would lead, through the messages conveyed by changes in prices, to adjustments across the entire economy.

Boettke (1997) traces the process by which Hayek, along with Ludwig von Mises, drew increasingly sharp distinctions between their thinking and the emerging Walrasian general equilibrium approach, partly in response to its effective use by Oskar Lange, Abba Lerner, and other proponents of the economic feasibility of central planning in the "socialist calculation" debates of the 1930s.[3] Lange, Lerner, and others argued that central planners could set prices and quantities to achieve the market outcome if they wished, but could also improve upon that outcome by taking into account externalities and other factors that a market would not consider. Hayek argued in response that it was impossible for central planners to choose prices and quantities that would achieve the market outcome, because the necessary information about preferences and production could not be known in advance, and only emerged through the process of market interaction.

Hayek's sharpest critique of the equilibrium model and the conception of competition on which it was built came in his 1948 paper "The Meaning of Competition." Here he argued that "the modern theory of competition deals almost exclusively with a state … in which it is assumed that the data for the different individuals are fully adjusted to each other, while the problem which requires explanation is the nature of the process by which the data are thus adjusted." That is, "the modern theory of competitive equilibrium *assumes* the situation to exist which a true explanation ought to account for as the effect of the competitive process."

In Hayek's (1948) view, assuming a state of equilibrium effectively precludes a serious analysis of competition, which he defines, following Samuel Johnson, as "the action of endeavoring to gain what another endeavors to gain at the same time." He continues as follows:

> Now, how many of the devices adopted in ordinary life to that end would still be open to a seller in a market in which so-called "perfect competition"

[2]The 1937 paper was originally read in 1936 as the presidential address to the London Economic Club. Glasner and Zimmerman (2014) note that the central arguments in this paper had been anticipated in a 1928 paper by Hayek in German.

[3]This debate led Hurwicz and others to develop the theory of mechanism design; see, for instance, Hurwicz's (1984) comment on Kirzner (1984). Maskin (2015) argues that two of Hayek's central claims—that the market mechanism is informationally efficient and incentive compatible—have been formally established in work by Mount and Reiter (1974), Jordan (1982), and Hammond (1979). These results show that the market mechanism is efficient at equilibrium without addressing whether an equilibrium is reachable. Furthermore, the mechanism uses prices that are centrally given, in that the same price vector is somehow transmitted to all market participants.

> prevails? I believe that the answer is exactly none. Advertising, undercutting, and improving ("differentiating") the goods or services produced are all excluded by definition—"perfect" competition means indeed the absence of all competitive activities.

He goes on to point out another absence in the standard model—social relationships among market participants:

> Especially remarkable in this connection is the explicit and complete exclusion from the theory of perfect competition of all personal relationships existing between the parties. In actual life, the fact that our inadequate knowledge of the available commodities or services is made up for by our experience with the persons or firms supplying them—that competition is in a large measure competition for reputation or good will—is one of the most important facts which enables us to solve our daily problems.

Is Hayek's Critique Obsolete?

Hayek's arguments have not been ignored by economists. Many of the important phenomena that cannot be accommodated by the Walrasian framework—advertising, undercutting, differentiating, reputation-building, and relational contracting—as well as other related phenomena such as bargaining and search, have been the focus of intense research effort over recent decades.[4] These advances explicitly allow for opportunistic and entrepreneurial behavior that goes well beyond the passive price-taking of agents in the Walrasian model, and this raises the question of whether Hayek's critique has been rendered obsolete by subsequent developments in the economics of information and applied game theory.

We think not. Economic analysis largely continues to be based on characterizations of equilibrium states, without attention to the processes through which such states might (or might not) be reached. For example, most contemporary models of strategic competition and search are equilibrium models, characterized by mutually consistent plans. These plans may have complicated features, with actions being contingent on history and the realization of random variables, but there is a common understanding across all individuals regarding the structure of the economy in which they are embedded. Left unaddressed is the process through which such a common understanding might arise.

[4]With the exception of the literature on mechanism design, discussed in the previous footnote, these developments have occurred quite independently of Hayek's thought. A notable exception is Makowski and Ostroy (2001), who argue that Hayek's critique of the standard model can be countered by reformulating that model with active rent-seeking agents, and redefining competitive equilibrium. Their proposed theory of markets takes explicit account of the concern that prices "will not be discovered unless opportunistic market participants find it in their self interest to reveal their trade-relevant private information."

This lack of attention to disequilibrium dynamics parallels the absence of an account of how a competitive equilibrium might arise in the Walrasian model itself. Hayek's critique of the latter applies also to richer conceptions of equilibrium in strategic settings with private and incomplete information. To see this point, consider Hayek (1948, p. 93):

> The problem becomes one of how the "data" of the different individuals on which they base their plans are adjusted to the objective facts of their environment (which includes the actions of the other people). Although in the solution of this type of problem we still must make use of our technique for rapidly working out the implications of a given set of data, we have now to deal not only with several separate sets of data of the different persons but also—and this is even more important—with a process which necessarily involves continuous changes in the data for the different individuals. … [T]he causal factor enters here in the form of the acquisition of new knowledge by the different individuals or of changes in their data brought about by the contacts between them.

Hayek's belief was that this process would lead to a diffusion of individually acquired knowledge across the economy and result in a more effective utilization of knowledge than would be possible under a centralized mechanism. In Hayek's view, the data that individuals have at their disposal consists of "abstract signals" including prices proposed, actions taken by others, and if bargaining actually takes place, information gained in the bargaining process even when no transaction was agreed upon (Kirman, Schulz, Härdle, and Werwatz 2005).

Most of the criticism that Hayek made of the various approaches to analyzing the functioning of the market process turned on the idea that the coordination of individual actions and beliefs is taken as given and the process by which this happens is not discussed. In his words (1948, p. 94):

> [T]he description of competitive equilibrium does not even attempt to say that, if we find such and such conditions, such and such consequences will follow, but confines itself to defining conditions in which its conclusions are already implicitly contained and which may conceivably exist but of which it does not tell us how they can ever be brought about. … competition is by its nature a dynamic process whose essential characteristics are abstracted away under the assumptions underlying equilibrium analysis.

Even within the Walrasian framework, the need to provide disequilibrium foundations for equilibrium analysis has been a recurring theme. Fisher (1983) was especially emphatic on this point, although he later wrote in a more pessimistic key: "The search for stability at great levels of generality is probably a hopeless one. That does not justify economists dealing only with equilibrium models and assuming the problem away" (Fisher 2011, p. 43).

When Prices Are Messages and Entrepreneurial Discovery is Destabilizing

While Hayek had little use for general equilibrium theory, he did implicitly assume that the process of entrepreneurial discovery would be stabilizing on average—that the profit opportunities that arose in disequilibrium would be exploited in a manner that sustained coherence and order in the system (Kirzner 1997). But the same problems of stability that have plagued general equilibrium theory also arise in the context of entrepreneurial discovery: individually profitable activities can be destabilizing in the aggregate.

In fact, the interpretation of prices as signals can itself give rise to destabilizing feedbacks, especially through the linkage of financial and goods markets. Because changes in asset prices can lead to substantial short-term capital gains and losses, information relevant to changes in such valuations will be actively sought. To the extent that a rise in the price of an asset can be used to infer that this happened as a result of the reaction of informed individuals to a change in the conditions of demand or supply, other individuals may seek to profit by buying and hoarding the asset in anticipation of further increases in price. But this activity itself has price effects, which in turn may result in rational hoarding by others, amplifying the destabilizing process.

To illustrate this problem, consider a classic passage from Hayek's celebrated 1945 paper:

> It is worth contemplating for a moment a very simple and commonplace instance of the action of the price system to see what precisely it accomplishes. Assume that somewhere in the world a new opportunity for the use of some raw material, say tin, has arisen, or that one of the sources of supply of tin has been eliminated. It does not matter for our purpose—and it is very significant that it does not matter—which of these two causes has made tin more scarce. All that the users of tin need to know is that some of the tin they used to consume is now more profitably employed elsewhere, and that in consequence they must economize tin (p. 526).

Not only do the agents not need to know much, according to Hayek, the process works well even if most of them know almost nothing. He continues:

> There is no need for the great majority of them even to know where the more urgent need has arisen, or in favor of what other needs they ought to husband the supply. If only some of them know directly of the new demand, and switch resources over to it, and if the people who are aware of the new gap thus created in turn fill it from still other sources, the effect will rapidly spread throughout the whole economic system and influence not only all the uses of tin, but also those of its substitutes and the substitutes of these substitutes, the supply of all the things made of tin, and their substitutes, and so on; and all this

> without the great majority of those instrumental in bringing about these substitutions knowing anything at all about the original cause of these changes.

The conclusion, Hayek reasons, is that:

> The whole acts as one market, not because any of its members survey the whole field, but because their limited individual fields of vision sufficiently overlap so that through many intermediaries the relevant information is communicated to all.

Suppose that the demand for tin has risen or the supply fallen, as Hayek postulates, and that the process he has in mind begins to operate. The price of tin begins to rise (though it cannot adjust instantaneously to the new equilibrium price). To an individual familiar with Hayek's argument, this change in price is *informative*: it is likely to have been caused by some changes in demand or supply. Recognizing this, such an individual may seek to profit by buying and hoarding tin in anticipation of further increases in price. But this activity itself has price effects, which in turn may result in hoarding by others, and so on. The changes in the price of tin will be driven by some combination of fundamental factors (of the kind that concern Hayek) and speculative forces that seek to extract information from prices. If speculative interest is strong enough, the result can be considerable nonfundamental volatility in the price of tin.

The mathematician Henri Poincaré recognized this problem as far back as 1908, after having been the examiner for Bachelier's (1900) pioneering thesis on market efficiency. Poincaré observed that the attempt to extract information from prices and other market signals could result in a form of herding that is not due to the psychological frailties of market participants, but arises simply because it makes economic sense in many instances to follow the crowd.

These effects can be captured by models of information cascades in which herding arises as a rational response to the extraction of information from the actions of others, as in the literature on observational learning (Banerjee 1992; Bikhchandani, Hirschleifer, and Welch 1992; Smith and Sorensen 2000). In this journal, Bikhchandani, Hirschleifer, and Welch (1998) survey this literature and explore the logic of this argument. And when there are strategic incentives to manipulate beliefs, information available to one party can be lost in the process of communication (Crawford and Sobel 1982).

In financial markets, attempts to extract information from prices can give rise to prolonged departures from fundamentals in theoretical models (Hong and Stein 1999; Abreu and Brunnermeier 2003), the empirical counterpart of which is excess volatility in prices (LeRoy and Porter 1981; Shiller 1981). When leverage is significant, relatively small informational shocks can give rise to large asset revaluations as funding dries up and assets must be liquidated at fire sale prices (Brunnermeier and Pedersen 2009; Adrian and Shin 2010; Geanakoplos 2010). Because information is costly to acquire and process, assets that have sufficient seniority are considered

safe under normal conditions; these can suddenly start to be perceived as risky and "information-sensitive" in crisis conditions, causing trading volume to collapse or markets to shut down entirely (Gorton 2012). Several of these mechanisms have been discussed by Brunnermeier (2009 in this journal) in the context of financial crisis of 2007–2008.

Such phenomena do not remain confined to the financial sector, because asset prices have real effects. One obvious example is the link between home values and new construction, but the point is considerably more general. The prices of claims on future income flows inevitably affect current production and consumption decisions, and prices of goods and services will not track relative resource scarcity consistently and reliably when assets are mispriced. And the most information-sensitive markets are subject to some of the most spectacular failures.

Hence, the economics of information does not lead us to a case for unregulated markets. But most of the above theory supporting this conclusion is obtained using equilibrium analysis, to which Hayek's many objections have been noted above. We next consider disequilibrium dynamics.

Disequilibrium Dynamics and Complex Adaptive Systems

The need to consider disequilibrium foundations of equilibrium economics has often been recognized, but explicit models of disequilibrium dynamics in economics remain rare. Exceptions include the work on learning in macroeconomics (Marcet and Sargent 1989; Woodford 1990; Evans and Honkapohja 2001). As in general equilibrium theory, general convergence results do not exist, although there are examples of sharp differences in the predictions of such models relative to those assuming equilibrium behavior throughout (Howitt 1992; García-Schmidt and Woodford 2015).

Further away from the mainstream there are models of complex adaptive systems, in which aggregate outcomes are determined by the social interaction of agents with limited and local knowledge. Epstein (2007) calls this approach to social science *generative*, while Tesfatsion (2006) calls it *constructive*. Its connection to Hayek's thought and method has been noted by Vriend (2002), Rosser (2012), and Axtell (2016), among others. This literature makes intensive use of computational rather than analytical methods, and it does not limit its focus to equilibrium paths; see Epstein and Axtell (1996) for an important and early contribution.

Among the earliest contributions to this literature is Schelling's (1971) model of segregation in self-forming neighborhoods. Here the agents are arrayed on a checkerboard grid. Each agent belongs to one of two groups. If agents are bordered by too great a proportion of neighbors from the other group, they will move. This process is repeated until a steady state is reached at which no agent wants to move. Schelling finds that integration can be sustainable once attained, but also that integrated states are extremely unlikely to be reached from arbitrary initial allocations, even when preferences are quite tolerant. That is, segregation is an *emergent*

property of the model, even though integration cannot theoretically be ruled out. It is not easy to obtain this insight through equilibrium analysis alone. And despite its simplicity, the model itself continues to be useful in organizing data (Cutler, Glaeser, and Vigdor 1999; Sethi and Somanathan 2004; Card, Mas, and Rothstein 2008; Bayer, Fang, and McMillan 2014).

Such agent-based models have also been successful in furthering our understanding of flows of pedestrian and vehicular traffic. Simple rules of avoidance can lead to flowing lines and other systematic patterns when density is low, but then as densities increase, bottlenecks, stop-go flows, and even gridlock can arise. Indeed, after the 2006 stampede in which close to 350 pilgrims died during the Hajj to Mecca, Dirk Helbing and his colleagues examined pedestrian crowd flows using computational methods in a collaboration with the Saudi government, and designed, implemented, and supervised a new set of pathways (Haase et al. 2016). The result was a substantial reduction in accidents.[5]

When this approach is applied to markets, then patterns of specialization, distribution, and prices arise as emergent properties of the interaction structure. That is, aggregate outcomes emerge that cannot be deduced analytically or in any other straightforward way from behavioral rules adopted by actors or any other attributes of individuals. A key element in this literature is the absence of *imposed* coordination across individuals in actions and beliefs. There is no assumption that individual plans are mutually consistent, or that subjectively perceived laws of motion coincide with the objectively realized laws of motion to which these perceptions give rise. There is no assumption that equilibrium markets clear, as in general equilibrium theory. This does not, of course, rule out model-consistent expectations or market clearing as *endogenous* outcomes, arising through responses by individuals.

A large and heterogeneous collection of models with these features is commonly grouped together under the umbrella of *agent-based computational economics*. The key components of the analysis are agents, which may be cognitively active units such as individuals, households, and firms, or inanimate components such as institutions for processing transactions or stocks of natural resources (Tesfatsion 2006). Agents may respond mechanically to inputs on the basis of physical laws or behavioral rules, or they may be sophisticated and forward-looking. They may be intertemporal optimizers employing the same dynamic programming methods used in orthodox models, but subject to private beliefs rather than mutually consistent expectations (Sinitskaya and Tesfatsion 2015). The key difference is that "events are driven solely by agent interactions once initial conditions have been specified. … [R]ather than focusing on the equilibrium states of a system, the idea is to watch and see if some sort of equilibrium develops over time" (Tesfatsion 2006).

Typically, agent-based models of financial markets involve a population of traders who make transactions based on their privately known and heterogeneous

[5]There is a tragic but informative postscript suggesting that the new system may have lacked resilience. In 2015, over 2,400 people were killed in a stampede on the Hajj, reportedly due to the closing of two of the five pedestrian routes to allow for the passage of important visitors invited by the royal family.

trading strategies. The payoffs to individual strategies are determined by these price dynamics, and successful strategies increase their presence in the population at the expense of less-successful ones. Such models have been able to replicate patterns in the data such as excess and clustered volatility, short-run momentum, and mean reversion over longer horizons. For surveys of how this approach has been used to understand patterns in asset price data, see LeBaron (2006) and Hommes (2006).

Leijonhufvud (2006) argued that agent-based process analysis "will finally make it possible to tackle the central problem of macroeconomics, namely, the self-regulating capabilities of a capitalist economy," but that the method remains in its "technical infancy." This assessment remains valid. Despite recent ambitious models of macroeconomic dynamics (Delli Gatti, Gaffeo, Gallegati, Giulioni, and Palestrini 2008; Sinitskaya and Tesfatsion 2015), financial fragility (Mandel, Landini, Gallegati, and Gintis 2015), and the housing bubble (Geanakoplos et al. 2012), there does not yet exist a canonical agent-based framework within which fundamental questions at the core of the discipline can be systematically explored.

The Verdict of the Market and the Verdict of History

The average size of firms in capitalist economies has been steadily increasing recently; indeed, there is a strong correlation between the average size of firms and income per capita. Gabaix (in this journal, 2016) argues that the increasingly skewed size distribution of US firms has led to some of these firms now becoming so important that changes in their performance can constitute major shocks to the macroeconomy. Given the vast scope of economic activity taking place within large firms, what this private and entirely apolitical discovery process reveals is the virtues of planning, albeit in private hands and subject to competitive forces.

Perhaps not surprisingly, Ronald Coase reported that much of the regular debate between himself and Hayek at the London School of Economics back in the 1930s centered on the subject of the firm as a centrally planned economy in miniature. In Coase's 1937 paper, he wrote that "the distinguishing mark of the firm is the suppression of the price system" in favor of a system in which a workman does what he does "because he is ordered to do so." Or more poetically, Coase quoted Dennis Robertson who said that firms were as "islands of conscious power in this sea of unconscious cooperation."[6]

But how could the suppression of the price system in favor of firm-based centralized planning possibly be a good thing? Kirzner (1992, p. 162) suggests a

[6]Herbert Simon (1991, p. 27) made the same point in this journal when he imagined "a mythical visitor from Mars" approaching earth in a spaceship "equipped with a telescope that reveals social structures. The firms reveal themselves, say, as solid green areas … Market transactions show as red lines connecting the firms, forming a network in the spaces between them. ... A message sent back home, describing the scene, would speak of 'large green areas interconnected by red lines.' It would not likely speak of 'a network of red lines connecting green spots.'"

reconciliation between Hayek's opposition to any form of planning and letting the market do the work when he says:

> In a free market, any advantages that may be derived from "central planning"… are purchased at the price of an enhanced knowledge problem. We may expect firms to spontaneously expand to the point where additional advantages of central planning are just offset by the incremental knowledge difficulties that stem from dispersed information.

In this version of Coasean thinking, market competition among firms will determine the appropriate extent of the market; the very process of entrepreneurial discovery that is the hallmark of Hayek's theory of competition is also the process that determines the boundary of the hierarchically organized firm. The verdict of the market, by this reasoning, substantially constrains the scope of activities that are conducted through markets rather than hierarchies.

Just as the verdict of the market constrains the sizes of individual firms, the verdict of history demarcates the boundary between state and market in the organization of economic activity. In *The Constitution of Liberty* Hayek argued that that "the value of freedom consists mainly in the opportunity that it provides for the growth of the undesigned, and the beneficial functioning of a free society rests largely on the existence of such freely grown institutions." By this logic, freely grown institutions that constrain the scope of the market in favor of public administration in resource allocation may be presumed to have purpose and value, even if these benefits cannot be deduced by rational reflection.

As it happens, most high-income countries have grown institutions that sharply constrain the operation of markets in many spheres, with the delivery of childhood education, health, and old-age pensions being prime examples. Economies with strong trade unions, large welfare states, and substantial regulation of the economy—all of which Hayek vociferously opposed—score well on measures of democracy, civil liberties, and innovativeness developed by the World Bank, Freedom House, and Bloomberg (World Bank 2017; Freedom House 2017; Jamrisko and Lu 2017). Indeed, the Nordic social democracies do slightly better by these measures, for example, than do the more *laissez faire* nations such as the United Kingdom and the United States.

The Road to Laissez Faire

Hayek believed that his economic vision provided the foundation for his support for free markets, but a careful reading of *The Road to Serfdom* (1944) suggests that he advocated minimal government involvement in economic activity because he saw hierarchical and collectivist political systems as a threat to individual liberty, not because his economics *per se* had demonstrated the superiority of unregulated

markets. The examples on his mind at the time—the Soviet Union under Stalin and Germany under Hitler—were convincing enough exhibits for his case. But, seven decades later, we have a record of sustained liberal democratic values in economies with substantial government involvement, and the evidence does not support Hayek's most dire predictions.

Fortunately, Hayek's economics and his political philosophy do not have to be taken as a package; it is possible to appreciate his insights into the functioning of a market economy without following him down the road to *laissez faire*. On this point we find ourselves agreeing with George Orwell (1944), who tempered an otherwise favorable evaluation of *The Road to Serfdom* with the caveat: "Professor Hayek … does not see, or will not admit, that a return to 'free' competition means for the great mass of people a tyranny probably worse, because more irresponsible, than that of the State."

We have not attempted here a comprehensive overview of Hayek's thought, which was extremely wide-ranging and has been ably summarized by others (see, for instance, Caldwell 2004). As noted by Glasner (1985): "Not, perhaps, since the Scottish Enlightenment philosophers for whom Hayek had such a strong affinity, has anyone made important contributions in a comparable range of disciplines." Hayek's vision of a decentralized solution to a massive and perpetually changing coordination problem involving autonomous entities will continue to shape the discipline well into the future.

Hayek dedicated *The Road to Serfdom* (1944) to "socialists of all parties" urging them to reconsider their understanding of the relationship between democracy and the organization of the economy. In a similar collegial spirit, we dedicate this modest effort to advocates of laissez faire inspired by Hayek, inviting them to reconsider what we have shown to be the tenuous link between Hayek's extraordinary contributions to economics and his opposition to any but the most minimal economic role for government.

■ *We thank Jeffrey Friedman, David Glasner, Gordon Hanson, and Timothy Taylor for their contributions to this essay and the Santa Fe Institute for providing an ideal environment for the collaboration that resulted in this paper.*

References

Abreu, Dilip, and Markus K. Brunnermeier. 2003. "Bubbles and Crashes." *Econometrica* 71(1): 173–204.

Adrian, Tobias, and Hyun Song Shin. 2010. "Liquidity and Leverage." *Journal of Financial Intermediation* 19(3): 418–37.

Axtell, Robert L. 2016. "Hayek Enriched by Complexity Enriched by Hayek." In *Advances in Austrian Economics: Revisiting Hayek's Political Economy*, edited by Peter J. Boettke and Virgil Henry Storr, 63–121. Bingley, UK: Emerald Insight.

Bachelier, Louis. 1900. "Théorie de la Spéculation." *Annales Scientifiques de l'E.N.S.* 17(3): 21–86.

Banerjee, Abhijit V. 1992. "A Simple Model of Herd Behavior." *Quarterly Journal of Economics* 107(3): 797–817.

Bayer, Patrick, Hanming Fang, and Robert McMillan. 2014. "Separate When Equal? Racial Inequality and Residential Segregation." *Journal of Urban Economics* 82: 32–48.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change in Informational Cascades." *Journal of Political Economy* 100(5): 992–1026.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1998. "Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades." *Journal of Economic Perspectives* 12(3): 151–70.

Boettke, Peter J. 1997. "Where Did Economics Go Wrong? Modern Economics as a Flight from Reality." *Critical Review* 11(1): 11–64.

Brunnermeier, Markus K. 2009. "Deciphering the Liquidity and Credit Crunch 2007–2008." *Journal of Economic Perspectives* 23(1): 77–100.

Brunnermeier, Markus K., and Lasse Heje Pedersen. 2009. "Market Liquidity and Funding Liquidity." *Review of Financial Studies* 22(6): 2201–38.

Caldwell, Bruce J. 2004. *Hayek's Challenge: An Intellectual Biography of F. A. Hayek*. University of Chicago Press.

Card, David, Alexandre Mas, and Jesse Rothstein. 2008. "Tipping and the Dynamics of Segregation." *Quarterly Journal of Economics* 123(1): 177–218.

Coase, Ronald H. 1937. "The Nature of the Firm." *Economica* 4(16): 386–405.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50(6): 1431–51.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. 1999. "The Rise and Decline of the American Ghetto." *Journal of Political Economy* 107(3): 455–506.

Debreu, Gérard. 1984. "La Supériorité de l'Économie Libérale Est Incontestable et Mathématiquement Démontrable." *Le Figaro,* March 10.

Delli Gatti, Domenico, Edoardo Gaffeo, Mauro Gallegati, Gianfranco Giulioni, and Antonio Palestrini. 2008. *Emergent Macroeconomics: An Agent-Based Approach to Business Fluctuations*. New York, NY: Springer.

Epstein, Joshua M. 2007. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press.

Epstein, Joshua M., and Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom up*. Washington, DC: Brookings Institution Press.

Evans, George W., and Seppo Honkapohja. 2001. *Learning and Expectations in Macroeconomics*. Princeton University Press.

Fisher, Franklin M. 1983. *Disequilibrium Foundations of Equilibrium Economics*. Cambridge University Press.

Fisher, Franklin. 2011. "The Stability of General Equilibrium—What Do We Know and Why Is It Important?" In *General Equilibrium Analysis: A Century after Walras*, edited by Pascal Bridel, 34–45. London, UK: Routledge.

Freedom House. 2017. *Freedom in the World 2017*. Washington, DC: Freedom House.

Gabaix, Xavier. 2016. "Power Laws in Economics: An Introduction." *Journal of Economic Perspectives* 30(1): 185–206.

García-Schmidt, Mariana, and Michael Woodford. 2015. "Are Low Interest Rates Deflationary? A Paradox of Perfect-Foresight Analysis." NBER Working Paper 21614.

Geanakoplos, John. 2010. "The Leverage Cycle." In *NBER Macroeconomics Annual 2009*, edited by Daron Acemoglu, Kenneth Rogoff, and Michael Woodford, 1–65. University of Chicago Press.

Geanakoplos, John, Robert Axtell, J. Doyne Farmer, Peter Howitt, Benjamin Conlee, Jonathan Goldstein, Matthew Hendrey, Nathan M. Palmer, and Chun-Yi Yang. 2012. "Getting at Systemic Risk via an Agent-Based Model of the Housing Market." *American Economic Review* 102(3): 53–58.

Glasner, David. 1985. "F. A. Hayek: Philosopher of the Open Society." *Michigan Quarterly Review* 24(4): 523–43.

Glasner, David, and Paul R. Zimmerman. 2014. "The Sraffa–Hayek Debate on the Natural Rate of Interest." Available at SSRN: http://ssrn.com/

abstract=2221695.

Gorton, Gary B. 2012. *Misunderstanding Financial Crises: Why We Don't See Them Coming.* Oxford, UK: Oxford University Press.

Haase, Knut, Habib Zain Al Abideen, Salim Al-Bosta, Mathias Kasper, Matthes Koch, Sven Muller, and Dirk Helbing. 2016. "Improving Pilgrim Safety during the Hajj: An Analytical and Operational Research Approach." *Interfaces* 46(1): 74–90.

Hammond, Peter J. 1979. "Straightforward Individual Incentive Compatibility in Large Economies." *Review of Economic Studies* 46(2): 263–82.

Hayek, Friedrich A. 1937. "Economics and Knowledge." *Economica* 4: 33–54.

Hayek, Friedrich A. 1944. *The Road to Serfdom.* University of Chicago Press.

Hayek, Friedrich A. 1945. "The Use of Knowledge in Society." *American Economic Review* 35(4): 519–30.

Hayek, Friedrich A. 1948. "The Meaning of Competition." Chap. 5 in *Individualism and Economic Order.* University of Chicago Press.

Hayek, Friedrich A. 1960. *The Constitution of Liberty.* University of Chicago Press.

Hayek, Friedrich A. 1968 [2002]. "Competition as a Discovery Procedure." *Quarterly Journal of Austrian Economics* 5(3): 9–23. (Article was in 2002; original lecture was in 1968.)

Hayek, Friedrich A. 1979. *Law, Legislation and Liberty: A New Statement of the Liberal Principles and Political Economy.* Vol 3: *The Political Order of a Free People.* University of Chicago Press.

Hayek, Friedrich A. 1994. *Hayek on Hayek: An Autobiographical Dialogue.* University of Chicago Press.

Hommes, Cars H. 2006. "Heterogeneous Agent Models in Economics and Finance." In *Handbook of Computational Economics*, vol. 2, edited by Leigh Tesfatsion and Kenneth L. Judd, 1109–86. Amsterdam: Elsevier.

Hong, Harrison, and Jeremy C. Stein. 1999. "A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets." *Journal of Finance* 54(6): 2143–84.

Howitt, Peter. 1992. "Interest Rate Control and Nonconvergence to Rational Expectations." *Journal of Political Economy* 100(4): 776–800.

Hurwicz, Leonid. 1984. "'Economic Planning and the Knowledge Problem': A Comment." *Cato Journal* 4(2): 419–25.

Jamrisko, Michelle, and Wei Lu. 2017. "These Are the World's Most Innovative Economies." *Bloomberg Markets*, January 17. https://www.bloomberg.com/news/articles/2017-01-17/sweden-gains-south-korea-reigns-as-world-s-most-innovative-economies.

Jordan, James S. 1982. "The Competitive Allocation Process Is Informationally Efficient Uniquely." *Journal of Economic Theory* 28(1): 1–18.

Kirman, Alan, Rainer Schulz, Wolfgang Härdle, and Axel Werwatz. 2005. "Transactions That Did Not Happen and Their Influence on Prices." *Journal of Economic Behavior and Organization* 56(4): 567–91.

Kirzner, Israel M. 1984. "Economic Planning and the Knowledge Problem." *Cato Journal* 4(2): 407–418.

Kirzner, Israel M. 1992. *The Meaning of Market Process: Essays in the Development of Modern Austrian Economics.* New York: Routledge.

Kirzner, Israel M. 1997. "Entrepreneurial Discovery and the Competitive Market Process: An Austrian Approach." *Journal of Economic Literature* 35(1): 60–85.

LeBaron, Blake. 2006. "Agent-Based Computational Finance." In *Handbook of Computational Economics*, vol. 2, edited by Leigh Tesfatsion and Kenneth Judd, 1187–1233. Amsterdam: Elsevier.

Leijonhufvud, Axel. 2006. "Agent-Based Macro." In *Handbook of Computational Economics*, vol. 2, edited by Leigh Tesfatsion and Kenneth Judd, 1625–37. Amsterdam: Elsevier.

LeRoy, Stephen F., and Richard D. Porter. 1981. "The Present-Value Relation: Tests Based on Implied Variance Bounds." *Econometrica* 49(3): 555–74.

Makowski, Louis, and Joseph M. Ostroy. 2001. "Perfect Competition and the Creativity of the Market." *Journal of Economic Literature* 39(2): 479–535.

Mandel, Antoine, Simone Landini, Mauro Gallegati, and Herbert Gintis. 2015. "Price Dynamics, Financial Fragility and Aggregate Volatility." *Journal of Economic Dynamics and Control* 51: 257–77.

Marcet, Albert, and Thomas J. Sargent. 1989. "Convergence of Least Squares Learning Mechanisms in Self-Referential Linear Stochastic Models." *Journal of Economic Theory* 48(2): 337–68.

Maskin, Eric S. 2015. "Friedrich von Hayek and Mechanism Design." *Review of Austrian Economics* 28(3): 247–52.

Mount, Kenneth, and Stanley Reiter. 1974. "The Informational Size of Message Spaces." *Journal of Economic Theory* 8(2): 161–92.

Nobelprize.org. 1974. Press Release, October 9. http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1974/press.html.

Orwell, George. 1944. "Review: The Road to Serfdom by F. A. Hayek/The Mirror of the Past by K. Zilliacus." *Observer* 9: 117–19.

Poincaré, Henri. 1908. *Science et Méthode.* Paris, France: E. Flammarion.

Rosser, J. Barkley, Jr. 2012. "Emergence and Complexity in Austrian Economics." *Journal of Economic Behavior and Organization* 81(1): 122–28.

Schelling, Thomas C. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology* 1(2): 143–86.

Sethi, Rajiv, and Rohini Somanathan. 2004. "Inequality and Segregation." *Journal of Political Economy* 112(6): 1296–321.

Shiller, Robert J. 1981. "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?" *American Economic Review* 71(3): 421–36.

Simon, Herbert A. 1991. "Organizations and Markets." *Journal of Economic Perspectives* 5(2): 25–44.

Sinitskaya, Ekaterina, and Leigh Tesfatsion. 2015. "Macroeconomies as Constructively Rational Games." *Journal of Economic Dynamics and Control* 61: 152–82.

Tesfatsion, Leigh. 2006. "Agent-Based Computational Economics: A Constructive Approach to Economic Theory." In *Handbook of Computational Economics*, vol. 2, edited by L. Tesfatsion and K. L. Judd, 831–80. Amsterdam: Elsevier.

Vriend, Nicolaas J. 2002. "Was Hayek an Ace?" *Southern Economic Journal* 68(4): 811–40.

White, Lawrence H. 1999. "Why Didn't Hayek Favor Laissez Faire in Banking?" *History of Political Economy* 31(4): 753–69.

Woodford, Michael. 1990. "Learning to Believe in Sunspots." *Econometrica* 58(2): 277–307.

World Bank. 2017. World Governance Indicators. http://databank.worldbank.org/data/databases/rule-of-law.

Journal of Economic Perspectives—Volume 31, Number 3—Summer 2017—Pages 231–238

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., St. Paul, MN 55105.

Potpourri

The IMF, World Bank, and World Trade Organization have combined to write "Making Trade an Engine of Growth for All: The Case for Trade and for Policies to Facilitate Adjustment." "According to simulation exercises, adjustment frictions in AEs [advanced economies] can lead to transition periods of up to 10 years and reduce the gains from trade by up to 30 percent … An unusual period of sharply increased import competition that began around 2000, along with other factors, appears to have negatively impacted regional labor markets in some AEs. Evidence on most episodes of trade increases suggests that the impact on aggregate labor market outcomes has been mild. When EMDEs [emerging market and developing

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.3.231

doi=10.1257/jep.31.3.231

economies] began to play a greater role in global manufacturing trade, in part reflecting the impact of pro-market reforms in China, a series of studies examined the impact on local labor markets during that period … These studies show that areas more exposed to competition from Chinese manufactures due to their industrial structure saw significant and persistent losses in jobs and earnings, falling most heavily on low-skilled workers." "When switching industries within manufacturing, workers in developed countries have been estimated to forego in terms of lifetime income the equivalent of 2.76 times their annual wage. Switching occupations may have similar costs, although these costs vary substantially across occupations and skill levels, with college-educated workers experiencing on average lower costs." "While employment protection legislation can reduce displacements, it can also impede the needed reallocation. There is broad consensus that employment protection should be limited, and that low hiring/firing costs coupled with protection through unemployment benefits is preferable, as in the case of Nordic countries … Well-designed and targeted trade-specific support programs can complement existing labor-market programs. … The effectiveness of these trade-specific programs has been mixed, however, and their coverage and size tends to be very small." March 22–23, 2017, http://www.imf.org/en/Publications/Policy-Papers/Issues/2017/04/08/making-trade-an-engine-of-growth-for-all.

Jason Furman delivered the Arnold C. Harberger Distinguished Lecture on Economic Development at the UCLA Burkle Center for International Relations, on "The Role of Economists in Economic Policymaking." He offers a wide range of concrete and useful examples. "I want to give you an example of a mistake I was involved in because I did not think hard about causation. At the end of 2008, I was working with Congress on legislation to raise the tax on tobacco products in order to pay for an expansion of the Children's Health Insurance Program (CHIP). The main proposal was to raise the tax on a pack of cigarettes from $0.39 per pack to $1.01 per pack. But we also needed to set tax rates on a wide range of other tobacco products including roll-your-own tobacco, pipe tobacco, small cigars, large cigars, and more. Amidst everything that was going on at the end of 2008 with the Great Recession I did not pay enough attention to this issue, even though I once sat through what felt like an endless meeting on the topic. What came out of that meeting was a proposal to raise the tax rate on roll-your-own tobacco by more than $20 a pound while leaving the tax rate on pipe tobacco largely unchanged. What followed was a huge decline in the sale of roll-your-own tobacco and a huge increase in the sale of pipe tobacco … It turns out that roll-your-own tobacco and pipe tobacco are highly substitutable—not because people have shifted to smoking pipes, but because you can still put pipe tobacco in a piece of paper, roll it up, and smoke it. This is not just a minor, technical observation. It turns out to be highly consequential for public health. I have estimated that the 2009 tobacco tax increase will reduce the number of premature deaths due to smoking by between 15,000 and 70,000 for each cohort. But it would have reduced them even more if we had harmonized the tax rate on different tobacco products, as we did in a subsequent proposal. In fact, economists in the Treasury Department estimated that the reduction in

tobacco consumption under a harmonization proposal would be nearly two and a half times the size it would be under an increase in the cigarette tax alone that raises comparable revenue." April 27, 2017. https://piie.com/system/files/documents/furman20170427.pdf.

Timothy J. Bartik has compiled "A New Panel Database on Business Incentives for Economic Development Offered by State and Local Governments in the United States." A short overview, "Better Incentives Data Can Inform both Research and Policy," appears in the Upjohn Institute *Employment Research* newsletter, and reports: "Using data from 1990 to 2015, the 'Panel Database on Incentives and Taxes' estimates marginal business taxes and business incentives for 45 industries in 33 states; the industries compose 91 percent of U.S. labor compensation, and the states produce over 92 percent of U.S. economic output. … Average incentives increased from 9 percent of business taxes in 1990 to 30 percent in 2015. … Because business executives tend to think in the short term, an incentive today is more effective at inducing location decisions than an incentive that is only paid out 10 years from now. The average state has incentives that are still 1.1 percent of business value-added when a facility is in its tenth year of operation. Reducing such long-term incentives would lower long-term government costs of incentives without having much effect on job creation. … Incentives designed as customized services may be more effective than tax incentives." Upjohn Institute. Full report from February 2017 is at http://research.upjohn.org/cgi/viewcontent.cgi?article=1228&context=reports. Newsletter (vol. 24, no. 2) is at http://research.upjohn.org/empl_research/vol24/iss2/1/.

Puzzling over Productivity

Edmund Phelps discusses "The Dynamism of Nations: Toward a Theory of Indigenous Innovation." "Some of the most serious faults of the once-dynamic economies lie in the private sector. A degree of corruption has seeped into some private institutions. The institution known as corporate governance is suspect. Most attempts at innovation are long-term projects shrouded in mystery, yet CEOs lean toward short-termism, aiming to maximize their bonuses and golden parachute by extracting every last gain in efficiency. … A characteristic of established and even accomplished corporations is that they are unable to go beyond a careful concern for efficiency, which demonstrates to the corporate board and shareowners their zeal. … [T]he rise of corporatism has transformed the functioning of the once-modern economies." "By now, corporatism is pervasive in all the nations of the West. Corporatism is behind the metastasis of vested interests, clientelism and cronyism that has brought a welter of regulations, grants, loans, guarantees, deductions, carve-outs, and evergreen patents mainly to serve vested interests, political clients, and cronies. In recent decades, large banks, large companies, and large government agencies formed a nexus to pump up home mortgage debt in America and to create unchecked sovereign debt and unfunded entitlements in several nations

in Europe. America has joined Europe in having a parallel economy that draws its nourishment from the ideas of political elites, whatever their motives, rather than from new commercial ideas. All this has combined to choke off much innovation." *Capitalism and Society*, 2017, vol, 2, no. 1, Article 3, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2963105.

Andrew G. Haldane delivered a lecture on "Productivity Puzzles" at the London School of Economics. "Growth theory would predict that, over time, technological diffusion should lead to catch-up between frontier and non-frontier countries. And the greater the distance to the frontier, the faster these rates of catch-up are likely to be. So what explains the 1¾ percentage point slowdown in global productivity growth since the 1970s—slower innovation at the frontier or slower diffusion to the periphery? If the frontier country is taken to be the United States, then slowing innovation can only account for a small fraction of the global slowing, not least because the US only has about a 20% weight in world GDP. In other words, the lion's share of the slowing in global productivity is the result of slower diffusion of innovation from frontier to non-frontier countries. … Taken at face value, these patterns are both striking and puzzling. Not only do they sit oddly with Classical growth theory. They are also at odds with the evidence of history, which has been that rates of technological diffusion have been rising rather than falling over time, and with secular trends in international flows of factors of production. At the very time we would have expected it to be firing on all cylinders, the technological diffusion engine globally has been misfiring. This adds to the productivity puzzle." March 20, 2017, http://www.bankofengland.co.uk/publications/Documents/speeches/2017/speech968.pdf.

Gustavo Adler, Romain Duval, Davide Furceri, Sinem Kiliç Çelik, Ksenia Koloskova, and Marcos Poplawski-Ribeiro have written "Gone with the Headwinds: Global Productivity." From the abstract: "[T]his note finds that the productivity slowdown reflects both crisis legacies and structural headwinds. In advanced economies, the global financial crisis has led to 'productivity hysteresis'—persistent productivity losses from a seemingly temporary shock. Behind this are balance sheet vulnerabilities, protracted weak demand and elevated uncertainty, which jointly triggered an adverse feedback loop of weak investment, weak productivity and bleak income prospects. Structural headwinds—already blowing before the crisis—include a waning ICT boom and slowing technology diffusion, partly reflecting an aging workforce, slowing global trade and weaker human capital accumulation." IMF Discussion Note, April 2017, SDN/17/04, https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2017/04/03/Gone-with-the-Headwinds-Global-Productivity-44758.

James Manyika, Jaana Remes, Jan Mischke, and Mekala Krishnan discuss "The Productivity Puzzle: A Closer Look at the United States." "We identify six characteristics that provide further insight into the productivity growth slowdown: declining value-added growth, a shift in employment toward lower productivity sectors, a relatively small number of sectors experiencing jumps in productivity, weak capital intensity growth across all types of capital, uneven rates of digitization across sectors (especially the large and often relatively

low-productivity ones), and slowing business dynamism." McKinsey Global Institute, March 2017, http://www.mckinsey.com/global-themes/employment-and-growth/new-insights-into-the-slowdown-in-us-productivity-growth.

Essays on Early Childhood Learning

Future of Children has devoted an issue to nine articles about "Social and Emotional Learning." From the introductory essay, "Social and Emotional Learning: Introducing the Issue," by Stephanie M. Jones and Emily J. Doolittle: "Research increasingly suggests that social and emotional learning (SEL) matters a great deal for important life outcomes like success in school, college entry and completion, and later earnings. This research also tells us that SEL can be taught and nurtured in schools so that students increase their ability to integrate thinking, emotions, and behavior in ways that lead to positive school and life outcomes. … All 50 states have SEL standards in place at the preschool level, and four (Illinois, Kansas, West Virginia, and Pennsylvania) have SEL standards for kindergarten through 12th grade. … At its core, SEL involves children's ability to learn about and manage their own emotions and interactions in ways that benefit themselves and others, and that help children and youth succeed in schooling, the workplace, relationships, and citizenship. … Decades' worth of research suggests that something other than academic skills and content knowledge strongly influences success in school and beyond. Indeed, SEL skills may be just as important as academic or purely cognitive skills for understanding how people succeed in school, college, and careers. In addition, preliminary evidence suggests that SEL skills could be central to understanding and remediating stubbornly persistent gaps in achievement defined by income and racial/ethnic differences. …" Spring 2017, http://www.futureofchildren.org/sites/futureofchildren/files/media/foc_spring_vol27_no1_for_web.pdf.

The Brookings Institution and the Duke University Center for Child and Family Policy convened a "Pre-Kindergarten Task Force of interdisciplinary scientists" to survey "The Current State of Scientific Knowledge on Pre-Kindergarten Effects," including Deborah A. Phillips, Mark W. Lipsey, Kenneth A. Dodge, Ron Haskins, Daphna Bassok, Margaret R. Burchinal, Greg J. Duncan, Mark Dynarski, Katherine A. Magnuson, and Christina Weiland. The report includes 10 short essays by specific authors, plus a "Consensus Statement," which says (in part): "Convincing evidence shows that children attending a diverse array of state and school district pre-k programs are more ready for school at the end of their pre-k year than children who do not attend pre-k. Improvements in academic areas such as literacy and numeracy are most common; the smaller number of studies of social-emotional and self-regulatory development generally show more modest improvements in those areas. Convincing evidence on the longer-term impacts of scaled-up pre-k programs on academic outcomes and school progress is sparse, precluding broad conclusions. The evidence that does exist often shows that pre-k-induced improvements in

learning are detectable during elementary school, but studies also reveal null or negative longer-term impacts for some programs." April 2017, https://www.brookings.edu/wp-content/uploads/2017/04/duke_prekstudy_final_4-4-17_hires.pdf.

Interviews with Economists

The Knowledge@Wharton website at the University of Pennsylvania has posted a 36-minute podcast interview with Angus Deaton, titled "Is Despair Killing the White Working Class? Ask Angus Deaton." "[I]f you look at white, non-Hispanics in midlife, in their early 50s for example, their mortality rate after 100 years of declining had turned the wrong way or at least flattened out. This is not happening to other groups in the U.S. It's not happening to Hispanics. It's not happening to African-Americans. And it's not happening in any other rich country in the world. This is happening to both men and women. Perhaps the most shocking thing is that a lot of the deaths come from what you might think of as behavioral factors, which are alcohol—alcoholic beverages—from suicides and from drug overdoses. Many of those drug overdoses are accidental overdoses from prescription drugs. People often think the health system is responsible for our health. In this case, the health system is responsible for killing people, not actually helping them. … There's a lot of really bad stuff going on, especially for this group without a B.A." April 6, 2017, http://knowledge.wharton.upenn.edu/article/despair-and-the-white-working-class.

Douglas Clement presents an "Interview with Gita Gopinath." As the subheading says, the main topics include the "dollar's unique status, crises & productivity, and policy spillover to emerging markets." "So what we analyzed in our paper is a set of fiscal instruments that would deliver the same outcomes as a currency devaluation. This idea goes back to Keynes, as you said, who proposed import tariffs and export subsidies as a substitute for currency devaluation. Given the illegality of using tariffs of this nature, we instead explored the role of value-added taxes and payroll subsidies or, more specifically, raising value-added taxes and cutting payroll taxes. What we found, surprisingly, is that this form of intervention did extremely well in mimicking the outcomes of a currency devaluation, not approximately but exactly. … Despite the virtues, there are political challenges to implementing a large fiscal devaluation. Countries live through a 10 percent exchange rate depreciation without immense anxiety, but if you raise value-added taxes by 10 percent, that would be very salient and likely politically infeasible. But the broader point we made was that there are instruments other than exchange rate devaluations that a country can use to gain trade competitiveness." *The Region*, Federal Reserve Bank of Minneapolis, December 20, 2016, https://www.minneapolisfed.org/publications/the-region/interview-with-gita-gopinath.

Aaron Steelman offers an "Interview" with Jonathan A. Parker. "[H]igh-income households used to live a relatively quiet life in the sense that the top 1 percent would earn a relatively stable income, more stable than the average

income. When the average income dropped by 1 percent, the incomes of the top 1 percent would drop by about only six-tenths of a percent. In the early 1980s that switched, so that in a recession if aggregate income dropped by 1 percent, the incomes of the top 1 percent dropped more like 2.5 percent—quadrupling the previous cyclicality. So now they're much more exposed to aggregate fluctuations than the typical income." "I use Nielsen Consumer Panel data to design and run my own survey on households to measure the effect of what was then the second of these large randomized experiments run by the U.S. government, the economic stimulus program of 2008. The key feature of that program was that the timing of the distribution of payments was determined by the last two digits of the Social Security number of the taxpayer, numbers that are essentially randomly assigned. So the government effectively ran a $100 billion natural experiment in 2008, distributing money randomly across time to people, and this policy provides a way to measure quite cleanly how people respond to infusions of liquidity. … The first thing I found out is that illiquidity is still a tremendous predictor of who spends more when a predictable payment arrives. … Low liquidity, or low financial wealth, is a very persistent state across households, suggesting the propensity to spend is not purely situational. A lot of it is closer to an individual-specific permanent effect than something transient due to temporary income shocks." *Econ Focus*, Federal Reserve Bank of Richmond, Third/Fourth Quarter 2016, pp. 22–26, https://www.richmondfed.org/publications/research/econ_focus/2016/q3-4/interview.

Cloud Yip offers a two-part interview with Ricardo Reis: "The Performance of Macroeconomics is Not that Bad!" and "Ricardo Reis Explains How to Use Interest on Reserve for Inflation Targeting." From the second part: "In the last six years, the world of central banking, the way central banks operate, the way they set monetary policy, has changed radically. … Reserves in the central banks used to be an asset that was essentially zero on the balance sheet. … Now it is one of the largest financial assets in the US. So, we have this new asset which is fundamental to the financial market, to the monetary policy, and it has fundamentally changed what the central bank balance sheet does. … A lot of my research in last year has been focused on understanding what does it mean and what does it imply for the control of inflation, for the risk of central bank insolvency and among others. That's what I called Reservism, trying to understand what is the role of this new asset called reserve has on the economy and the central bank policy. … Reserves right now are overnight deposit in central bank by banks, they are paid a given interest rate but once you started thinking about what they are, you realized that those could be different. They could, instead of promising an interest rate, promising a different payment. They could be, instead of overnight, a 30-day deposit. They could be lots of different things." EconReporter (an independent Hong Kong journalism project). Part 1 of the interview, posted February 9, 2017, is at http://en.econreporter.com/en/2017/02/ricardo-reis-performance-macroeconomics-not-bad. Part 2, posted February 11, 2017, is at http://en.econreporter.com/en/2017/02/ricardo-reis-explains-central-banks-can-use-interest-reserve-target-inflation.

Discussion Starters

Nicholas Bloom discusses "Corporations in the Age of Inequality." "The real engine fueling rising income inequality is 'firm inequality': In an increasingly winner-take-all or at least winner-take-most economy, the best-educated and most-skilled employees cluster inside the most successful companies, their incomes rising dramatically compared with those of outsiders. This corporate segregation is accelerated by the relentless outsourcing and automation of noncore activities and by growing investment in technology." *Harvard Business Review*, March 2017, https://hbr.org/cover-story/2017/03/corporations-in-the-age-of-inequality.

Daniel Griswold takes on the task of "Plumbing America's Balance of Trade." "America's commerce with the rest of the world must be and always is balanced when taking into account investment flows as well as the exchange of goods and services. … [O]ne key insight for public policy is that the total outflow of dollars each year from the United States to the rest of the world is matched by an equal inflow of dollars from the rest of the world to the United States. There is no need to worry about a 'leakage' of dollars siphoning off demand from the domestic economy. Dollars spent on imported goods and services return to the United States, if not to buy US goods and services, then to buy US assets in the form of an inward flow of investment." Mercatus Center at George Mason University, March 2017, https://www.mercatus.org/system/files/mercatus-griswold-balance-of-trade-v1.pdf/.

Jacob Udell and Glenn Yago discuss "Still Digging Out: The Economics of a Palestinian Future." "Though the labor force participation rate [in the West Bank and Gaza] is currently at its highest since 2000 (at an unimpressive 46 percent), it has been accompanied by an overall spike in unemployment—implying that the net entry of job seekers into the market exceeds the ability of the economy to create employment. Meanwhile, the Palestinian Authority has also become the employer of last resort, with 23 percent of the workforce on its rolls. The wave of youth entering the labor market in the past decade, coupled with the frictional and structural unemployment of the adult population and almost nonexistent job growth, has left youth unemployment at alarming levels. Since 2001, for instance, unemployment among males aged 15–24, which seems to function as a leading indicator of civil unrest, has averaged 35 to 40 percent and reached 43 percent in 2014. Since 2005, real average wages have decreased by some 10 percent, while unemployment remains at around one-quarter of the labor force, and average GDP growth lags behind population growth by 2.6 percent per year. And while considerable sums flow into the territories from overseas Palestinians, there are no 'diaspora bonds' or other vehicles to facilitate investment by Palestinian ex-pats (whose wealth estimates by the World Bank have varied from $40 billion to $80 billion). One mark of a lack of confidence in the economy: Palestinian investment abroad in 2015 was $5.9 billion—$1.3 billion more than foreign investment in Palestine." *Milken Institute Review*, Second Quarter 2017, pp. 78–85. http://www.milkenreview.org/articles/still-digging-out.

The *JOE Network* fully automates the hiring process for the annual economics job market cycle.

For:

JOB CANDIDATES

- Search and Save Jobs
- Create a Custom Profile
- Manage Your CV and Applications
- Get the Attention of Hiring Committees
- Apply for Multiple Jobs from One Site
- Request Reference Letters

EMPLOYERS

- Post and Manage Job Openings
- Search Candidate Profiles
- Manage Applications and Materials
- Collect Reference Letters
- Download Applicant Data
- Share Candidate Materials

FACULTY

- Manage Letter Requests
- Upload Custom or Default Letters
- Track Task Completion Status
- Assign Surrogate Access
- Minimize Time Investment

This hiring season, take advantage of the AEA's enhanced JOE Network targeted to the comprehensive needs of all participants in the annual economics job market cycle.

The ***JOE Network*** automates the hiring process. Users share materials, communicate confidentially, and take advantage of new features to easily manage their files and personal data. Everything is securely maintained and activated in one location. The JOE Network is accessible right from your desktop at the AEA website.

Experience the same great results with more features, more time savings, and a beginning-to-end process.

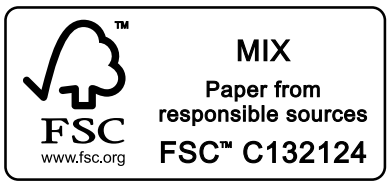
AMERICAN ECONOMIC ASSOCIATION

Try the JOE Network today!

www.aeaweb.org/JOE

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary *JEP* for your e-reader, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $20.00, $30.00, or $40.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.



The Journal of

Economic Perspectives

Summer 2017, Volume 31, Number 3

AMERICAN ECONOMIC ASSOCIATION