The Journal of *Economic Perspectives*

A journal of the American Economic Association

Spring 2015

The Journal of
Economic Perspectives

A journal of the American Economic Association

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

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

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

No responsibility for the views expressed by the authors in this journal is assumed by the editors or by the American Economic Association.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Spring 2015, Vol. 29, No. 2. 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 29 • Number 2 • Spring 2015

Symposia

The Bailouts of 2007–2009

Disability Insurance

Articles

Features

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

A Retrospective Look at Rescuing and Restructuring General Motors and Chrysler[†]

Austan D. Goolsbee and Alan B. Krueger

The rescue of the US automobile industry amid the 2008–2009 recession and financial crisis was a consequential, controversial, and difficult decision made at a fraught moment for the US economy. Both of us were involved in the decision process at the time, but since have moved back to academia. More than five years have passed since the bailout began, and it is timely to look back at this unusual episode of economic policymaking to consider what we got right, what we got wrong, and why.

We are pleased and a bit surprised by how well the last five years have played out for the domestic auto industry. At a critical point in the internal debate over the auto industry bailouts in March 2009, Larry Summers, at that time director of the National Economic Council, assembled members of the Obama administration's economic and autos team around his cramped table in the West Wing of the White House. He held a straw vote on whether the advisors believed Chrysler would survive for five years if a government-supported merger with Fiat went through. A narrow majority, including us, voted no. Five years on, both General Motors and Chrysler have survived, rebounded, and, by many metrics, appear healthy.

■ *Austan D. Goolsbee is Robert P. Gwinn Professor of Economics, University of Chicago, Chicago, Illinois. He was a member of the Council of Economic Advisers from March 2009 to September 2010, and Chairman of the Council from September 2010 to August 2011. Alan B. Krueger is Bendheim Professor of Economics and Public Affairs, Princeton University, Princeton, New Jersey. He was Assistant Secretary for Economic Policy and Chief Economist at the US Treasury Department from 2009 to 2010, and Chairman of the Council of Economic Advisers from November 2011 to August 2013. Their email addresses are goolsbee@chicagobooth.edu and akrueger@princeton.edu.*

[†] To access the Data Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.29.2.3

doi=10.1257/jep.29.2.3

Economic analysis contributed throughout the process of deciding how to respond to the auto companies' requests for extraordinary support, and President Obama weighed the economic arguments as well as the political and social realities. We agreed with others in the administration that it was essential to rescue General Motors to prevent an uncontrolled bankruptcy and the failure of countless suppliers, with potentially systemic effects that could sink the entire auto industry. Our analysis suggested that a failure of the much smaller Chrysler, however, would probably not have systemic effects for the whole industry and that rescuing the company would make it more difficult and more costly for taxpayers to rescue GM, although we recognized that a failure of Chrysler would cause considerable hardship to its workers and their families and communities. In the end, the president made the decision to rescue both General Motors and Chrysler and to put them through a tough restructuring via bankruptcy.

It is hard to argue that this decision did not deliver important economic benefits to the recovery and country, although the government did not recover the full amount of TARP funds it invested. If GM and Chrysler had been allowed to fail, in all likelihood the Great Recession would have been deeper and longer, and the recovery that began in mid-2009 would have been weaker. The rescue has been more successful than almost anyone predicted at the time. Some of this success resulted from actions the auto companies took; some happened because the rebound in consumer demand for autos has been especially strong during the last five years. The auto industry has turned out to be one of the drivers of the economic recovery. Yet we suspect that the conditions that led the auto bailout to be a success were fairly unique in American economic history, and, we hope, unlikely to be repeated anytime soon.

In this article, we describe the events that brought two of the largest industrial companies in the world to seek a bailout from the US government, the analysis that was used to evaluate the decision (including what the alternatives were and whether a rescue would even work), the steps that were taken to rescue and restructure General Motors and Chrysler, and the performance of the US auto industry since the bailout. We close with some of the general lessons to be learned from the episode.

How the US Auto Industry Imploded

In the run-up to the 2009 bailout, the "Big Three" US automakers recorded some of the worst corporate performances in American history. General Motors alone lost almost $40 billion in 2007 and another $31 billion in 2008. Ford lost $3 billion and then $15 billion. Chrysler was a privately held company that did not disclose earnings publicly, but was losing comparable amounts of money. The Great Recession that began in late 2007 had a catastrophic impact on the automakers. Auto sales plummeted in 2008 and again in 2009 to below 10 million, from a peak of more than 17 million just a few years earlier.

By fall 2008, the financial situation of the domestic automakers was so dire that they would soon be unable to make their wage and supplier payments. In November 2008, the chief executive officers of Ford, General Motors, and Chrysler came before the House and the Senate to request a $25 billion working capital "bridge loan" from the US government to enable them to make these payments and to help keep them out of bankruptcy and avoid possible liquidation. In the auto executives' view, the crisis they were facing centered on macroeconomic forces outside of their control. Chrysler CEO Robert Nardelli (2008) explained at the outset of the hearing, "We are asking for assistance for one reason: To address the devastating automotive industry recession caused by our Nation's financial meltdown." He said that buyers' and dealers' lack of access to credit was preventing them from buying vehicles and wrecking the automakers' business. They were asking for capital to tide them over, with no conditions attached, until the economy returned to normal so that they could avoid bankruptcy or liquidation.

Of course, no one knew if the 17 million annual sales rates achieved earlier in the 2000s would ever return. Auto credit had been unsustainably inflated by the same housing and credit bubble that led to the economic crisis in 2008. The ratio of cars-to-population and the fraction of auto buyers stretching their credit by using subprime auto loans were both at record highs. If demand rebounded only partway toward its previous high after the recession ended, it was not clear that all of the "Big Three" automakers could survive.

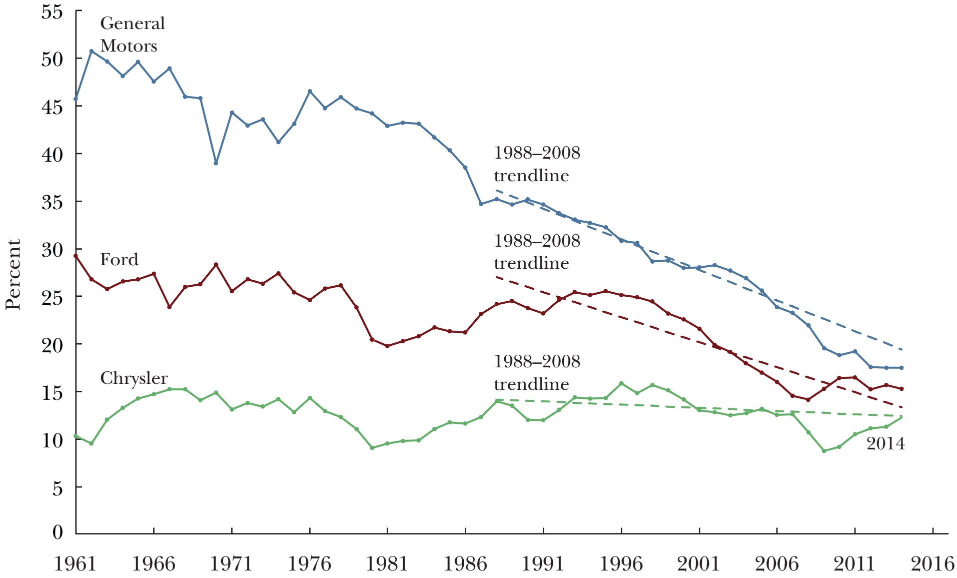
When critics highlighted the US auto industry's decades-old problems of high cost, questionable quality, and the like as factors contributing to the industry's troubles during the financial crisis, the executives argued that they had already done the restructuring necessary to fix those problems, so that they were no longer an issue. In reality, the Big Three automakers' problems had built up over many years and were certainly not solely a result of the economic downturn.

Falling demand was a persistent and severe problem for the Big Three. Market share trends weighed heavily against them. Figure 1 plots the US market share of each of the Big Three automakers in the decades running up to the crisis as a percentage of total auto sales. There was a sustained and substantial downward trend in demand of more than 2 percentage points per year for the Big Three combined. The Big Three's share in 1998 was 71 percent; by 2008, it was 47 percent. These negative trends were especially severe for GM, the largest of the domestic companies.

If anything, these declines in market share understate the severity of the dwindling demand facing the manufacturers. The Big Three had been engaged in substantial price discounting relative to the competition. By 2008, the Big Three were discounting comparable cars by $2,000 to $3,000 (Helper 2010). A number of factors had taken a toll on the demand for cars from the Big Three manufacturers over time: the widespread perception of perennial quality and reliability issues, lower resale values, poorly received new models, and a lack of low-gas-mileage cars at times of rising fuel costs.

Moreover, the "transplant" car factories—that is, domestic US production of foreign-owned companies like Honda, Toyota, Nissan, and others—were expanding

Figure 1
"Big Three" Automakers' Shares of US Total Vehicle Sales



employment and production in the United States using predominantly nonunion plants in the American South, even as the Big Three automakers struggled. For example, from 2000 to 2013, employment at the domestic transplant carmakers almost doubled to 163,000, while Big Three employment fell steadily and was cut nearly in half to 253,000, according to *Automotive News* data reported in Kurylko (2013). This pattern suggested that the problems of the Big Three legacy US automakers were perhaps particular to those firms, not to the national automobile manufacturing industry.

A common refrain among industry analysts and critics in Congress was that US automakers were uncompetitive versus their foreign counterparts as well as against the transplant factories. Estimates of the hourly compensation of the Big Three automakers put hourly compensation almost 25 percent higher than in the transplants (Leonhardt 2008). After including the legacy costs of retirees, average labor costs for the Big Three were almost 45 percent higher. In addition, a surprisingly large share of labor compensation for the Big Three automakers was a fixed cost, rather than a variable one. Pension and health care costs for retirees are obvious fixed costs, but the United Automobile Workers (UAW) had also negotiated for workers to be paid 95 percent of their salary when they were on layoff, which in effect turned mostly variable labor compensation into a fixed cost. Under these conditions, it was hard to see how a rescue could make the Big Three more cost competitive with rivals at home and abroad for more than a short time, unless it reduced the fixed costs associated with retirees, the uncompetitive compensation levels for existing workers, and the crushing interest payments owed to bondholders.

To summarize, the problems facing the automakers included long-term falling market share, compounded by a massive short-term drop in aggregate demand, with large fixed costs. This combination resulted in huge short-term losses. But even if the automakers could reduce their fixed costs and even if the recession ended and aggregate demand returned to normal levels in the short-run, unless they could stop their persistent decline in market share, these automakers would soon be back in trouble.

By December 2008, regardless of what one thought the sources of the Big Three's problems were or what should or should not have been done in the preceding years, General Motors and Chrysler faced an existential threat. Congress could not agree to provide the automakers emergency financing and adjourned for the holidays at the end of 2008, leaving the Big Three scrambling. The Bush administration decided to tap into Troubled Asset Relief Program (TARP) funds authorized under the Emergency Economic Stabilization Act (that had been signed into law on October 3, 2008). It lent GM and Chrysler more than $20 billion to keep them afloat into early 2009. Of that amount, $17.5 billion went directly to the automakers. The rest went to the financing arms of these firms, the General Motors Acceptance Corporation (GMAC) and Chrysler Financial. Ford decided not to take government support. Ford had large losses but had borrowed a significant amount of money in 2006 and begun restructuring before the financial crisis struck, so the company was able to withstand the cash crunch.

GM and Chrysler received these loans with the condition that they develop plans to make themselves "viable" as ongoing enterprises. The firms were given until February 2009 to come up with the plans. The Obama administration took office in late January.

The viability plans that the companies submitted in February 2009 were summarily rejected as unrealistic and inadequate, which sent the effort back to the drawing board. The gap in time between the granting of the loans in December 2008 and agreement on a workable plan for restructuring the companies and making them financial viable meant that the interim $20 billion in loans made to keep the companies afloat while they prepared the original viability plans was unlikely ever to be repaid.

A first obvious consideration was whether General Motors and Chrysler could just enter one of the standard paths for companies in dire financial trouble. For example, one common approach is for the troubled firm to borrow funds using so-called "Debtor-in-Possession" financing. This new source of financing is allowed to be senior (that is, it would be paid first) to all existing company debt. In the meantime, a distressed company can sell off key pieces to acquire cash, perhaps on the way to finding a full buyer in the intermediate term. But in early 2009, these options were merely fantasy. The financial crisis raged. To be sure, there were speculations early in 2009 that perhaps a large Chinese or other national sovereign wealth fund would be willing to buy major portions of the companies but there was, realistically, no chance of these outcomes happening in the requisite timeframe—if they ever would have happened at all. Even if such a buyer had materialized, scrutiny of these kinds of transactions by antitrust authorities, along with the Congress and its

Committee on Foreign Investment in the United States, would have taken months and faced a high chance of falling through. There was speculation about a merger of GM and Chrysler, but it was unclear that a merger of two failing companies would solve either of their problems.

Later, during the presidential election of 2012, critics of the rescue argued that private lenders should have been allowed to fund the General Motors and Chrysler restructurings in bankruptcy. In early 2009, however, such funding simply did not exist. At that moment, for better or for worse, it was government money or bust. Without government funds, GM and Chrysler were on a path to disorderly bankruptcy, which, by all accounts, would take years for resolving the myriad disputes among thousands of creditors, suppliers, and so on, and would likely mean liquidation.

The Costs of Not Rescuing

What were some of the more likely outcomes if the government had not acted in early 2009 to extend further assistance to GM and Chrysler? As we and others in the Obama administration investigated this question, the answers we heard were not comforting. The companies themselves would lay off their workers immediately. There would be widespread spillovers into supplier industries and auto dealerships, as well as knock-on macroeconomic effects through a *reverse* multiplier. The Congressional Oversight Panel (2009) called the companies' possible collapse "a potentially crippling blow to the American economy that Treasury estimated would eliminate nearly 1.1 million jobs."[1] Other contemporary estimates suggested that the near-term jobs at risk from a disorderly liquidation could reach as high as 2.5 to 3.3 million jobs (Zandi 2008; Cole et al. 2008; Scott 2008).

It was easy to question the methodology of some of the more extreme job loss estimates. For example, although we believe that a bankruptcy reorganization of GM and Chrysler under Chapter 11 would have been so disorderly as to be economically wasteful and destructive, presumably some proportion of the assets of the firms would have been put to use. However, we felt confident that a collapse of both companies would have resulted in the immediate loss of at least 500,000 to 1 million jobs. Total job losses from a messy liquidation of Chrysler by itself, in our estimation at the time, would have been in the neighborhood of 300,000 jobs.

Setting aside the costs to the individuals involved, we knew that job losses of this scale would impose sizable costs on various levels of government through the need for additional spending on safety net, health care, unemployment insurance, and other programs, and we sought to quantify some of these costs. In addition, because the company pension funds would probably also be bankrupted, tens of billions of dollars in pension liabilities would be transferred to the Pension Benefit Guaranty Corporation, which was itself already in a precarious financial position. In

[1] Actually, the original job estimates came from the Council of Economic Advisers under Edward Lazear rather than Treasury.

considering the costs and benefits of a rescue plan for GM and Chrysler, one had to acknowledge that the alternative of letting the companies proceed into a disorganized bankruptcy would not be "free."

Of course, this is not to say that government should try to keep all large companies alive because their failure would be painful. We certainly had no desire to put the US economy on the path we perceived that Japan had followed in the preceding decades, where stagnation had continued for years as the government propped up "zombie firms" that were not viable companies. Further, the auto industry is highly capital-intensive compared with other industries, so if one measures jobs saved on a bang-for-the-buck basis, using money to support other industries might have a larger employment impact. Moreover, public opinion polling suggested that large majorities opposed bailouts for any firms, including auto companies.

As the policy team grappled with these issues, a consensus emerged that allowing *both* companies into uncontrolled bankruptcy was ill-advised. We heard numerous experts opine that a failure of General Motors, in particular, would level a major blow to supply chains and to consumer confidence that would have an outsized negative impact on spending as well as the argument that this was the equivalent of negative stimulus precisely when the fiscal and monetary policy authorities were attempting to provide positive stimulus. The negative aggregate impact of a disorderly failure of GM would be too great at exactly the wrong moment for the economy. Thus, the question arose of whether we should rescue GM but let Chrysler, the smaller and weaker of the two firms, go into a disorderly bankruptcy.

We had several concerns about the merits of a Chrysler bailout. First, auto sales had plummeted from 16.5 million units in 2006 to 9.5 million in 2009. Our forecasts at the time, and those of many industry analysts, suggested that US auto sales in the steady state would be around 15 to 15.5 million a year. We thought that Chrysler and GM, which had been losing market share for decades, were viable restructured businesses if the market was over 16 million cars, but would there be sufficient demand for *both* Chrysler and GM to be profitable in the long run? Trying to keep each of the Big Three in operation with such a low rate of sales might endanger them all.

Second, our internal research and reading of the industrial organization literature on demand elasticities in the auto industry indicated that consumers who buy from Chrysler would likely turn to Ford or GM if their preferred Chrysler model was not available. Table 1 illustrates this point with sales data from 2008 by market segment. About 75 percent of Chrysler's sales were concentrated in large cars, minivans, SUVs, and trucks. This was almost double the share of sales in those segments in the full passenger vehicle market. Non-Chrysler demand in those segments was heavily domestic: two-thirds of non-Chrysler sales in these Chrysler-heavy segments went to GM or Ford. Even these numbers understate the degree of overlap among the domestic firms by not including minivans and full-size pickup trucks such as the Toyota Sienna, Honda Odyssey, or Toyota Tundra that were not made by the Big Three, but were still domestically produced in the transplant factories. If consumer demand starts with choosing a segment (that is, the kind of car you wish to buy like a minivan or a sports car) and then a particular model, Chrysler's failure might have

Table 1
Sales by Market Segment

Segment	*Share of total Chrysler sales*	*Share of total market sales*	*GM + FORD share of non-Chrysler sales*
Full-size pickup	22.2	12.3	87.4
Minivan	21.5	4.5	11.7
Mid-size SUV	10.5	10.1	48.3
Full-size SUV	9.6	4.8	77.6
Full-size	8.8	5.5	83.1
Sports car	1.9	1.8	66.7
	74.5	**39.0**	**65.8**
Compact	12.3	18.8	30.3
Mid-size	7.1	16.4	23.2
Compact SUV	3.3	7.9	42
Mid-size pickup	2.4	2.5	22.3
	25.1	**45.6**	**29.3**
Entry luxury	0	4.1	0
Subcompact	0	2.9	14.7
Mid-size luxury SUV	0	2.5	13.7
Mid-size luxury	0	2	34.9
Full-size luxury	0	1	45.7
Full-size luxury SUV	0	1	70
Compact pickup	0	0.5	98.1
Sports car luxury	0	0.5	1.8
MPV	0	0.4	0
Compact luxury SUV	0	0.3	0
	0	**15.2**	**20.5**
TOTAL	**100**	**100**	**41.4**

Note: The model-level sales data were compiled by *Automotive News*, and we obtained them from the *Good Car Bad Car* archives at http://www.goodcarbadcar.net/2013/02/2008-america-auto-sales-rankings-by-model.html, and then summed them by the segment definitions in the Wikipedia Car Classification page.

a much smaller impact on the economy than people feared. Chrysler's failure might, for example, simply mean that Dodge Ram buyers would, instead, buy another full-sized pickup, and all of those models are produced domestically. Nationwide net employment loss from Chrysler's liquidation in this type of situation would be much smaller than the national estimates suggested, as consumers would switch to other domestically produced cars in the absence of Chrysler. Also, letting Chrysler fail would have substantially reduced the amount of money needed to rescue GM and would have increased the profitability outlook for GM and Ford.

Third, Chrysler had been acquired and restructured twice before without success. The merger between Daimler-Benz and Chrysler that took place in 1998, but was dissolved in 2007, had proved unsuccessful in a more favorable economic environment. The buyout by private equity company Cerberus in 2007 had been unable to stem the problems and instead added more years of malaise and mismanagement. We saw little prospect that a purchase of Chrysler by Fiat would provide more synergies or a more reassuring brand name for American consumers. Furthermore, automobiles are a business with large economies of scale and Chrysler operated at a notably smaller scale than the largest car companies like GM, Toyota, Ford, and others—even with Fiat as a partner.

From a hard-nosed triage view, it was unclear why Chrysler should receive special treatment, especially given that public bailout money could probably save more jobs in a less-capital-intensive industry and a liquidation of Chrysler did not seem to pose a systemic threat. Even if our fears were accurate that the failure of Chrysler would cause 300,000 workers employed there and in the auto supply chain to lose their jobs (assuming no substitution to other domestic producers in the short run), the US labor market in early 2009 was in miserable shape. Job separations at this time were running at 4 to 5 million per month in the private sector workforce according to data from the Job Opportunities and Labor Turnover Survey (JOLTS), and net job losses at this time (after hiring was taken into account) were running around 700,000 per month. Indiscriminate carnage from the financial crisis existed in virtually *every* industry, not just the auto industry.

Of course, there were also economic arguments in favor of rescuing Chrysler. First, although we expected that shortfalls in supply caused by the failure of Chrysler could in time be picked up by an expansion of the other auto manufacturers, and that viable segments of Chrysler's business—such as its minivan unit or Jeep division—would eventually be acquired by other auto companies, "eventually" could take a long time. A messy liquidation of Chrysler would make the transition costs higher.

Another important factor in the decision related to the nature of the auto industry itself, which threatened a kind of negative contagion because of company interdependence. Over the preceding decades, a larger and larger fraction of the value-added in the auto industry had migrated to auto suppliers. Large suppliers of seats, electrical systems, and other components normally supplied multiple car companies, and many of the largest auto suppliers such as Lear, American Axle, and Visteon were in dire financial shape. Hundreds of suppliers were known to be teetering on the edge (Stoll and McCracken 2009; Kiley 2009; Helper 2010).

The Motor and Equipment Manufacturers Association (2009) submitted data showing that 66 percent of Chrysler suppliers were also suppliers to GM and 54 percent were suppliers to Ford. In previous years, even some seemingly modest supplier disruptions or specific parts shortages resulting from strikes or natural disasters had caused widespread disruption to the production lines of car manufacturers. If auto suppliers failed because of lost demand from a Chrysler liquidation, it could easily disrupt the other US producers, both in Detroit and in the transplant firms elsewhere. Ford itself was arguing, publicly, for their competitors GM and Chrysler to receive bailouts on the grounds that their failure would endanger Ford's own production. We feared a chain reaction.

As the academic legal debate over bankruptcy law has observed, bankruptcy is largely a micro solution, aimed at reorganizing the assets and liabilities of a single firm (Warren 1987; Baird 1987). It is not a macro solution. It does not take cross-industry spillovers or broader government or social costs into account. The auto taskforce attempted to quantify and weigh many of these factors, though there was much disagreement on the details and magnitudes. For example, our early estimates of job losses and supplier impacts often came from the industry's own representatives, who had an incentive to exaggerate their estimates. One of our roles, for instance, was to note that about half of the employees in the auto supply chain were involved in manufacturing replacement parts, which still would have been in demand even with a failure of domestic automakers.

The Decision and the Aftermath

President Obama heard the analysis on all sides of the issue. He concluded that the economy should not risk the failure of both companies in 2009 and opted to rescue both General Motors and Chrysler. Rattner (2010, p. 120) notes, "The case for saving Chrysler was based more on political and social reality." President Obama made the decision to reject the viability plans the companies submitted from the first round of loans in February 2009 and ordered a new and more serious restructuring effort, led by a team of private sector turnaround experts that he brought into the administration. Separate from the efforts made to reorganize the car manufacturers, the rescue effort also included providing money to the affiliated finance companies and auto suppliers, and guaranteeing warranties to customers.[2]

In an industry with high fixed costs, annual profitability is largely determined by total market demand—known in the auto trade as the Seasonally Adjusted Annual Rate (SAAR) of lightweight vehicle sales—along with market share and price. Price depends on perceived quality and resale value. We examine developments in costs, product quality, prices, market share, and SAAR below.

Massive Restructuring and Cost Reduction

We knew that a lasting restructuring of General Motors and Chrysler would likely require a number of steps: reducing their legacy costs (payments to bondholders and retirees), reducing their number of dealers, cutting capacity and weaker brands, and expanding a two-tier structure where newly hired workers were paid less than incumbents. In March 2009, President Obama instructed his auto team, "I want you to be tough and I want you to be commercial" in regards to setting terms for an alliance between Chrysler and Fiat and restructuring GM (Rattner 2010, p. 132). The funds that the US Treasury provided to Chrysler and GM came with strict requirements on their restructuring. Because of their different financial positions, most of the support provided to GM took the form of equity, while support for Chrysler was

[2] A clever market-based mechanism was used to extend credit to critical suppliers by giving automakers access to funds to use to keep their critical suppliers afloat. However, only $413 million of $5 billion allocated to this program was lent to suppliers; all of it was eventually repaid to Treasury.

in the form of debt that needed to be repaid. One could justify the less-generous terms of support for Chrysler in part because Chrysler was in more precarious financial shape than GM in 2009, and in part because Chrysler was less-pivotal for the near-term course of the auto industry and economy given its smaller size.

As a condition of the earlier government loans, General Motors agreed to cut its debt by $30 billion by converting existing debt into equity. It also agreed to cut employment from 96,000 to 45,000 by 2012; bring its labor costs in line with those of the transplants by 2012; sell its Saab, Saturn, and Hummer divisions; and reduce its number of models from 45 to 40. GM failed to meet the full conditions of the bailout, and its chief executive officer, Rick Wagoner, was replaced in March 2009. On June 1, 2009, GM filed for bankruptcy with $173 billion in liabilities and $82 billion in assets. The company closed a dozen plants and eliminated more than 20,000 jobs. Stockholders were wiped out and bondholders were issued new stock worth much less than the value of their bonds. More than 1,100 of 6,100 dealerships would eventually close. GM emerged from bankruptcy quickly, on July 10, 2009, as two separate companies. About half of the members of the board of directors were replaced, and several top executives were dismissed or reassigned. The old company retained the liabilities, and a "Shiny New GM" held the assets and soon became profitable, earning its first annual profit in ten years in 2010. Retiree health benefits, funded by an entity known as a voluntary beneficiary benefits association (VEBA), were cut for GM's more than 330,000 retirees and surviving spouses in the United States, and the VEBA was funded primarily with an equity stake in the company.

Chrysler filed for bankruptcy on April 30, 2009. The company closed 789 of its 3,200 dealerships as part of its bankruptcy reorganization. More than a dozen plants closed. Under agreement with the United Autoworkers union, the two-tier wage system was expanded, with wages for new hires cut to about half of the $29 per hour that longtime union members earned (although these wages were then raised to $17 an hour in 2011). Defined benefit pensions were eliminated for new hires and replaced with 401(k) plans. Overall wage and benefit costs at Chrysler and GM were brought down to be roughly in line with those at Honda and Toyota plants operating in the United States. Benefits provided by Chrysler's voluntary beneficiary benefits association (VEBA) were also slashed, and the VEBA received a 55 percent equity stake in the company. Fiat gained minority ownership and corporate control of the restructured Chrysler.

Restructuring the two failing auto companies reduced their fixed and variable costs at the expense of much pain for their creditors, workers, managers, and dealers.[3] Just as importantly for their long-run success, the new management of the companies sought to improve the culture of their organizations and introduce better business

[3] Anticipating that restructuring the companies would cause much pain and disruption, we recommended that the President establish a Director of Recovery for Auto Communities and Workers to coordinate agencies and resources across the federal government to ease the transition for hard-hit communities and workers. Economist Edward Montgomery, now at Georgetown University, ably served in this capacity.

practices to produce higher-quality cars. From brakes, wheels, and suspension to styling and advertising—including popular commercials featuring Eminem and Clint Eastwood launched during the 2011 and 2012 Super Bowls—an attempt was made to improve the culture and quality of work at Chrysler, in particular. Chrysler posted a profit in the first quarter of 2010. When asked what had changed at Chrysler, Fiat chief executive officer Sergio Marchionne (2014) recently responded: "The culture; the technology that's in place; the way in which the cars are manufactured; the attitude of the workforce; the efficiency; the land speeds; the output of the system has completely changed. I mean, if you took a Japanese guy into our plant today he'd be impressed." Marchionne also offered a simple explanation for why Chrysler was able to change so quickly: "I know that when you're broke you change your ways a lot faster."

Price Discounts and Perceived Quality

In the longer term, we knew that for the auto companies to survive they needed also to deal with the falling demand for their products. Prior to the financial crisis, General Motors and Chrysler concentrated on producing larger, less-fuel-efficient, and more-costly-to-produce models than their competitors, and offered aggressive price discounts to consumers.

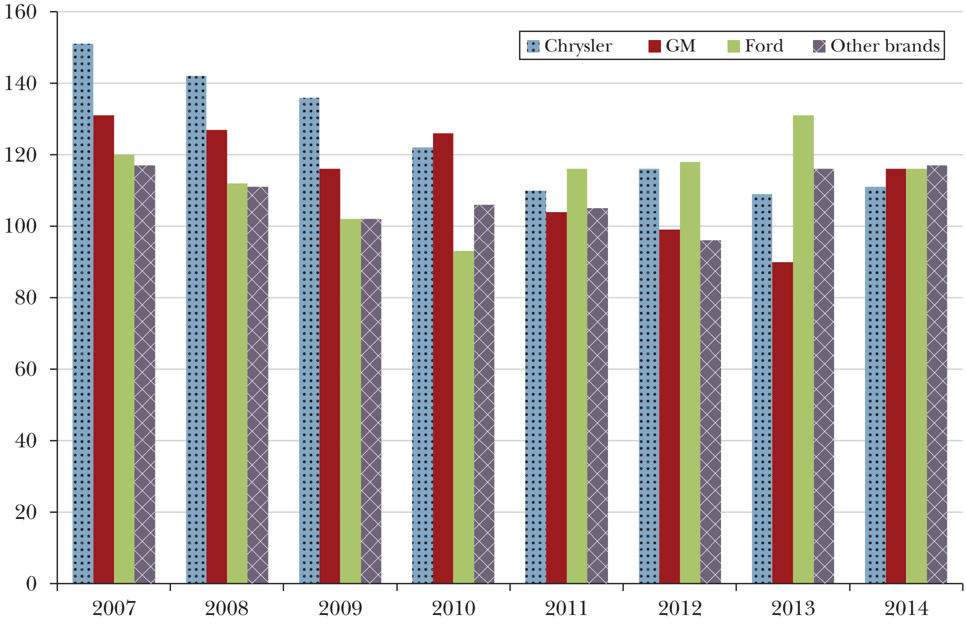
Since the restructuring, there are some signs that quality has improved and that price discounting has become less aggressive, though the jury is still out. Figure 2 reports the JD Power quality rating for Chrysler, GM, and Ford, and for all other automakers combined. JD Power's Initial Quality Study provides information on new-vehicle quality based on a survey of a nationally representative sample of car buyers (results weighted to reflect sales). The questionnaire asks car owners to indicate which, if any, problems they have experienced from a list of 228 possible items, and they can write in any additional problems not included on the list. Figure 2 reports the number of problems per 100 vehicles. A lower figure indicates fewer problems and higher quality. Although this measure is crude (one reason is that some problems are worse than others), it is a common metric of quality in the industry.

In 2010 and earlier years, owners of new General Motors and Chrysler vehicles reported a higher incidence of problems than owners of other cars. Starting in 2011, however, this measure of quality improved considerably for both firms, with the number of problems reported per new car about on par with that of the other auto manufacturers.

However, in 2014 General Motors agreed to pay the US Department of Transportation the maximum civil penalty of $35 million for failing to report and delaying a recall of 2003–2011 cars with defective ignition switches and airbags that failed to deploy, a problem that GM reportedly was aware of at least as early as November 2009. In total, GM recalled 29 million cars in North America as of the middle of 2014, breaking the record for most recalled cars in any full year. Chrysler has launched its own recalls for ignition switches. These recalls point to clear quality problems. Overall, the extent to which quality has improved since 2010 remains an open question.

Data that allow quality-adjusted price comparisons among cars are sketchy, but indicate that the Detroit brands continued to offer steeper discounts than

Figure 2
JD Power Quality Rating
(problems per 100 vehicles)



Notes: Figure 2 reports the JD Power quality rating for Chrysler, GM, and Ford, and for all other automakers combined. The rating is based on the JD Power's Initial Quality Study, which provides information on new-vehicle quality from a survey of a nationally representative sample of car buyers (results weighted to reflect sales).

other automakers after 2009; nonetheless, their discounts and incentives relative to the industry average fell by about 10 percent from 2002–2008 to 2009–2011. Chrysler's CEO Sergio Marchionne, in particular, has waged a campaign against price discounting, emphasizing, "Unprofitable volume is not volume I want." He reportedly berated Chrysler's head of sales, who was dismissed shortly afterwards, for seeking to offer price rebates along with "Cash for Clunkers," the colloquial name for the Car Allowance Rebate System that the federal government operated in July and August 2009 to give people an incentive to trade in their older cars for more fuel-efficient models (Linebaugh and Bennett 2010). General Motors had reduced its sales incentives below those of Chrysler and Ford by February 2014, but the company subsequently sharply increased discounts to counteract a drop in demand due to adverse publicity over the recalls in spring 2014 (Kessler and Vlasic 2014).

Market Share

The market share of each of the Big Three automakers was presented earlier in Figure 1. As a benchmark, the graph also shows the trend projected from a linear regression over the period 1988–2008. General Motors' market share has been on a downward trajectory for the past 50 years, falling from 50.7 percent of the market in 1962 to 40.4 percent in 1985, 30.6 percent in 1997, and 19.6 percent in 2009. Ford's market share has also trended down from 29 percent in 1961 to 14 percent in 2008, with a notable reversal in the period from 1981 to 1995, and then a sharper decline through 2008. Chrysler's market share, by contrast, fluctuated between 10 and 15 percent from 1961 to 2008, and fell to an all-time low of 8.8 percent in 2009.

After 2009, Chrysler's share of the market rose for five consecutive years, its best performance since the early 1990s. Chrysler's market share stood at 12.3 percent in the first half of 2014, which was 3.5 percentage points, or 40 percent, above its 2009 level. These gains have been widely attributed to the improved management and higher-quality product initiated by Sergio Marchionne. The drop in gasoline prices at this time also probably boosted the Big Three's market shares above what they otherwise would have been by raising demand for larger vehicles.

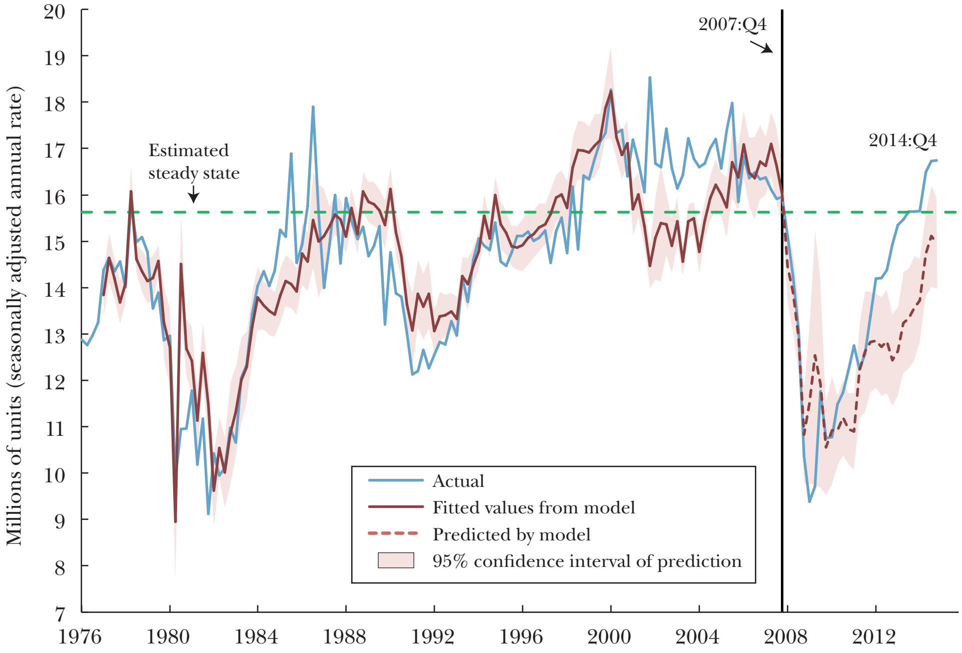
One of our main concerns about the auto rescue was that the domestic brands to a considerable degree compete with each other, and so rescuing Chrysler, the weakest and smallest of the three firms, would make it harder (and more expensive for taxpayers) for General Motors to survive. There appears to be some support for this view, as GM's market share continued to decline after 2009, and its decline was at least as quick as it was over the preceding two decades. The fact that GM eliminated four unprofitable brands—Saturn, Pontiac, Hummer, and Saab—also undoubtedly contributed to its decline in market share after 2009.

It is impossible to know what would have happened to GM's market shares had Chrysler been liquidated in 2009, but the data in Figure 1 show a notably strong rebound in Chrysler's market share, from a historically low base, and a continuation of GM's decades of long decline. The market share of the Big Three combined stood at 45.1 percent in the first half of 2014, above their 2009 combined low of 43.7 percent in 2009, but well below their share of 50.5 percent on the eve of the economic crisis in 2007. These figures suggest that, to some extent, Chrysler's gains did come at the expense of the other domestic firms.

Rebound in Aggregate Auto Demand

The biggest factor contributing to the positive recovery of the automakers, however, has been the rapid rebound of consumer demand for autos more generally. Auto sales are normally procyclical. Figure 3 shows auto sales each quarter since 1976. We see that nationwide sales plummeted during the Great Recession, falling to their lowest quarterly level since the deep 1981 recession. Many factors affect car sales, in addition to the state of the economy, such as population growth, credit availability, and the age and durability of the existing fleet. We and many industry analysts expected sales to bounce back to around 15 to 15.5 million a year when the economy normalized. In its submission to the government in February 2009, GM's baseline forecast of annual sales was 16 million units in 2012 and market share of 20 percent. (GM was too optimistic: in 2012, actual sales were 14.4 million and GM's market share was just 17.6 percent.)

Figure 3
Lightweight Vehicle Sales: Actual and Modeled Results
(millions of units; seasonally adjusted annual rate)



Note: The figure shows fitted values from a regression model to predict lightweight vehicle sales (the “seasonally adjusted annual rate” or SAAR) for 1976–2007, and the projected values from 2008 forward. (See Table 2 for the regression results).

To compare actual sales to what one would predict from a forecasting model, we regressed quarterly sales of lightweight vehicles (adjusted to the “seasonally adjusted annual rate” or SAAR) on real GDP growth, the unemployment rate, population growth, the Federal Reserve’s Senior Loan Officers’ Survey (SLOOS) measure of willingness to lend to consumers, the logarithm of the average real price of a gallon of gasoline in the previous quarter, and the standard deviation of gas prices over the preceding four quarters, using a sample from 1977:Q1 to 2007:Q4. (The sample begins in 1977 because gasoline price data from the Energy Information Administration are available starting in 1976.) The regression results are presented in Table 2, and Figure 3 shows the fitted values during the sample period and the projected values from 2008 forward. The explanatory variables account for 72 percent of the variability in quarterly car sales.

Most of the coefficients associated with the explanatory variables have their expected signs. For example, sales are stronger when the economy is stronger (that is, faster GDP growth or lower unemployment) and when credit conditions are looser. Higher gas prices are associated with lower sales, although the relationship is weak and statistically insignificant. Greater variability in gas prices, however, is

Table 2

Regression Model to Predict Lightweight Vehicle Sales, 1977–2007

(quarterly sales adjusted to the "seasonally adjusted annual rate" or SAAR)

	Mean of variable (standard deviation)	*Coefficient (standard error)*
Real GDP Growth (%)	3.20 (3.10)	0.099 (0.025)
Unemployment Rate (%)	6.12 (1.41)	−1.150 (0.093)
Population Growth (%)	1.28 (0.58)	0.226 (0.116)
SLOOS credit availability	109.24 (16.57)	0.044 (0.009)
log gasoline price (lagged)	0.63 (0.24)	−0.027 (0.808)
Standard deviation of log gasoline price over previous four quarters	0.056 (0.036)	8.657 (4.033)
Constant	—	15.948 (1.357)
R^2		0.716

Notes: We present results from a regression of quarterly sales of lightweight vehicles (the "seasonally adjusted annual rate" or SAAR) on real GDP growth, the unemployment rate, population growth, the Federal Reserve's Senior Loan Officers' Survey (SLOOS) measure of willingness to lend to consumers, the logarithm of the average real price of a gallon of gasoline in the previous quarter, and the standard deviation of gas prices over the preceding four quarters, using a sample of 124 quarterly observations from 1977:Q1 to 2007:Q4. Real GDP Growth, Unemployment Rate, and Population Growth are seasonally adjusted and at an annual rate. The log of the real price of gasoline is for the previous quarter, and gas prices were deflated by the Personal Consumption Expenditures deflator. The standard deviation of log real gas prices is computed over the preceding four quarters. The mean (standard deviation) of the dependent variable is 14.8 million (2.0 million) SAAR. In the second column, in parentheses, are Newey-West standard errors with four lags.

associated with higher sales, as households may adjust their model of car in response to recent movements in gas prices.

The model effectively captures the collapse in auto sales during the Great Recession, and predicts most of the rebound since the recession officially ended in mid-2009, although it underpredicts actual sales in 2012–14 (see Figure 3). In the last quarter 2014, actual sales were 1.8 million above the level the model would predict at a seasonally adjusted annual rate. Part of the rebound in car sales appears to represent overshooting of actual sales relative to the prediction of the simple model. This pattern is not wholly unexpected given the pent-up demand that accumulated during the Great Recession, and the fact that the parsimonious regression model used here ignores dynamics. There was some significant overshooting of sales early in two of the three previous recoveries as well.

In early 2009, the respected economic forecasting firm Macroeconomic Advisers, which had expected a strong economic recovery (GDP growth of 3.9 percent and unemployment rate of 5.8 percent in 2013), predicted that auto sales would reach 15.4 million in 2013. The fact that auto sales slightly exceeded that amount at 15.5 million, despite their overly optimistic assumptions about the state of the economy, is a sign that the rebound in auto sales exceeded expectations given the actual path of the recovery.

To gauge the importance of the rebound in sales for the fate of the auto rescue, suppose that domestic auto sales had remained at 9.5 million instead of rebounding to 16.5 million in 2014:Q2. In this scenario, Chrysler would have needed to raise its market share by 12.4 percentage points to achieve the actual volume of sales it registered in 2014:Q2. Thus, Chrysler's impressive 3.5 percentage point gain in market share was far less significant than the overall rebound in market demand.

We can use the coefficients from the regression model in Table 2 to derive an estimate of "steady state" car sales. Specifically, we assumed the values of the explanatory variables would equal the forecast of real GDP growth and unemployment used by the Obama administration for the "out year" forecasts in 2023, which are best understood as an estimate of long-run underlying trends. Specifically, we assume a 2.3 percent rate of GDP growth and an unemployment rate of 5.4 percent, which correspond to the 2023 forecasts in the administration's FY2015 Budget (Table 2-1). We assume a growth rate for the civilian non-institutional population of 0.9 percent, corresponding to the 2023 baseline forecast in CBO's February 2014 "Budget and Economic Outlook." For the SLOOS credit availability variable, log of real gas prices, and standard deviation of log gas prices, we use the average values over the period 2002:Q1 to 2007:Q4. This calculation suggests that steady state annual car sales will be around 15.6 million.

If our estimate of steady state car sales is correct, sales may slip by about 7 percent from their current level. For Chrysler, this amounts to about a quarter of their post-restructuring gain in market share. Given the restructuring of costs, we suspect that there will be sufficient demand to sustain the Big Three at their current level of market share. In addition, there is room for GM potentially to raise its profitability by implementing some of the tough measures that Chrysler has implemented. But steady state market demand is probably just large enough to sustain the existing domestic firms, although there is little margin for the companies to be viable ongoing concerns if they are mismanaged in the future.

Autos and Industrial Recovery

Even in the information age, the auto industry remains a major contributor to the US economy. Moreover, modern automobiles are advanced manufacturing products. We were told by Ford, for example, that the value of electronics, software, and intellectual property accounts for about 30 percent of the average vehicle's price.

Manufacturing played a critical role in the recovery from the Great Recession, and autos played an outsized role in the manufacturing recovery. Five years after the start of the recovery, the rise in motor vehicles and parts production accounted for more than 25 percent of the rise in total manufacturing industrial production, even though motor vehicles and parts account for only about 6 percent of total manufacturing value added. Although it is not unusual for the auto industry to punch above its weight early in a recovery, it has played an unusually large role relatively long into the current recovery. At the same point in the last four recoveries, motor vehicles and parts accounted for only 11 percent of the rise in manufacturing production, on average.

Since bottoming at 623,300 jobs at the trough of the recession in June 2009, employment in the motor vehicles and parts manufacturing industry has increased by 256,000 jobs (as of July 2014). This is a stark contrast from the previous recovery, when jobs in the industry steadily declined. The increase in the number of jobs in motor vehicles and parts manufacturing accounted for nearly 60 percent of the total rise in manufacturing jobs in the recovery's first five years. In addition, some 225,000 jobs have been added at motor vehicle and parts dealers. Counting both manufacturers and dealers, auto-related jobs accounted for 6 percent of the total 8.1 million jobs that were added, on net, in the first five years of the recovery—triple the sector's 2 percent share of total employment. Although the auto sector played an outsized role in the recovery, it should also be apparent that given the relatively low share of total employment in autos and related jobs, there is a limit to how much the auto rebound could have driven a jobs recovery.

Exit Strategy

The US Treasury Department provided roughly $80 billion in assistance to the auto industry: $51 billion to GM, $12.5 billion to Chrysler, and $17.2 billion to what is now Ally Financial, but was formerly GMAC Finance (US Department of the Treasury 2015). By the end of 2014, the government had closed all three of these positions.

At the urging of Larry Summers, the Obama administration established principles for its role as majority owner of General Motors. These included: setting upfront business goals and selecting executives and a strong board of directors; only voting as a shareholder on major corporate governance issues or major transactions; letting the board and management run the company; and selling the government's shares as soon as practical to recover taxpayer money and return the company to private ownership. A similar approach was taken to Chrysler. From the outset, we were determined to avoid the problem that had worsened Japan's stagnation in the 1990s and 2000s of propping up zombie companies for long periods of time when they should have ceased to exist. As President Obama (2009) put it, his goal was "to get GM back on its feet, take a hands-off approach and get out quickly."

On December 9, 2013—much sooner than virtually anyone expected—the government fully exited its investment in General Motors by selling its remaining shares, and critics could no longer say that GM stood for Government Motors. The

US Treasury recovered a total of $39.7 billion from its investment of $51.0 billion in GM. By the end of 2014, Treasury sold its remaining stake in Ally Financial, recovering $19.6 billion from the original $17.2 billion investment in Ally, for a $2.4 billion gain for taxpayers. In May 2011, Chrysler repaid its outstanding loans from the Troubled Asset Relief Program (TARP) six years ahead of schedule. Chrysler returned $11.2 billion of the $12.5 billion it received through principal repayments, interest, and cancelled commitments, and the Treasury fully exited its connection with Chrysler. In January 2014, Fiat purchased the shares in Chrysler owned by the voluntary employee's benefits association (VEBA) that funded retiree health benefits and took full ownership of Chrysler.

For the most part, the Obama administration adhered to its goals and avoided political meddling. There were some notable exceptions, however. For example, when GM's Chief Executive wanted to move the company's headquarters from the Renaissance Center in Detroit to its Tech Center in Warren, Michigan, to be closer to the workforce—which made some business sense—the administration blocked the move. Congress and the administration both set restrictions on executive compensation for companies that had received Troubled Asset Relief Program funds (for example, the annual compensation for chief executive officers was capped at $9.5 million). The administration included a "vitality commitment" as a condition of receiving funding, which prevented the companies from moving work at US plants to other countries. Members of Congress frequently attempted to intervene to prevent unnecessary and inefficient dealerships from being closed, to the administration's consternation.

Some have argued that the rescue improperly paid unsecured union workers ahead of unsecured bondholders due to political pressures. The wider debate about what is permitted and encouraged by bankruptcy law and how those rules might have applied to this specific rescue situation is beyond our scope, but we have a few observations. First, as a legal matter, a large majority of bondholders voted for the deal and a bankruptcy judge approved it. That is why it proceeded. The agreement was not unilaterally imposed by the Obama administration. Second, there were legitimate business reasons why one might need to pay some unsecured creditors so the firms would be able to continue operating. Guaranteeing the warrantees of car owners, for example, also prioritized unsecured creditors. But if consumers did not trust the warranties, demand for cars likely would drop precipitously. Likewise, if workers refused to accept the deal or shirked on their duties, the automakers' viability as an ongoing concern was in jeopardy. Similar payments were made to workers in the bankruptcies of the steel companies in the 1980s, where there was not a government rescue. Third, despite their haircut, bondholders almost certainly received well more than they would have under the alternative scenario in which the government did not intervene in the depths of the crisis. Finally, despite insinuations to the contrary, incumbent workers took dramatic cuts to their benefits and bore substantial risk when the voluntary beneficiary benefits association (VEBA) that funded retiree health benefits for a time held a substantial equity share of the firms.

Conclusion

Economists and economic analysis had a key seat at the table in the decision to rescue and restructure General Motors and Chrysler. The decision was risky. Those of us involved gathered all the information we could find and tried to put, finally, the companies on a sustainable footing. We did not know if it would work. In particular, we had reservations about the long-run viability of the Chrysler–Fiat merger. In an interview in the *Detroit News* (Shepard 2015), President Obama explained his decision this way: "There was clear-eyed recognition that we couldn't sustain business as usual. That's what made this successful. If it had been just about putting more money in without restructuring these companies, we would have seen perhaps some of the bleeding slowed but we wouldn't have cured the patient."

To their credit, the two companies restructured to a greater degree than they had ever done before and under extreme pressure, and—after shedding much legacy debt—returned to profitability in 2010. They also were fortunate that the economy began to turn around and that consumer demand for autos rebounded strongly.

It is fair to say that no one involved in the decision to rescue and restructure General Motors and Chrysler ever wanted to be in the position of bailing out failed companies or having the government own a majority stake in a major private company. We are both thrilled and relieved with the result: the automakers got back on their feet, which helped the recovery of the US economy. Indeed, the auto industry's outsized contribution to the economic recovery has been one of the unexpected consequences of the government intervention. The automakers' future success will depend on their own managerial decisions in the years to come. The fact that Ford was able to weather the economic downturn and financial crisis because it had taken precautionary steps and efforts to restructure before calamity hit, while GM and Chrysler could not have survived without extraordinary government support, is a stark reminder of the importance of good managerial decisions for the survival of businesses.

■ *The authors are grateful to Anthony Casey, Steve Rattner, Harry Wilson, Timothy Taylor, David Autor, and Chang-Tai Hsieh for helpful comments and to David Cho and Loullya Saney for excellent research assistance.*

References

Baird, Douglas. 1987. "Loss Distribution, Forum Shopping, and Bankruptcy: A Reply to Warren." *University of Chicago Law Review* 54: 815.

Cole, David, Sean McAlinden, Kristin Dziczek, and Debra Maranger. 2008. "CAR Research Memorandum: The Impact on the U.S. Economy of a Major Contraction of the Big Three Automakers." Center for Automotive Research, Nov. 4.

Congressional Oversight Panel. 2009. "The Use of TARP Funds in the Support and Reorganization of the Domestic Automotive Industry." September Oversight Report. www.gpo.gov/fdsys/pkg/CHRG-111shrg51964/html/CHRG-111shrg51964.htm.

Helper, Susan. 2010. "Managing the 2020 Auto Supply Chain: Developments to Watch." Presentation at the Chicago Fed conference "After the Perfect Storm: What's Next for the Auto Industry," May 10. www.chicagofed.org/digital_assets/others/events/2010/automotive_perfect_storm/helper.pdf.

Kessler, Aaron M., and Bill Vlasic. 2014."To Lift Sales, G.M. Turns to Discounts." *New York Times*, July 31.

Kiley, David. 2009. "Billions for Auto Suppliers' Bailout." *Bloomberg BusinessWeek*, March 19.

Kurylko, Diana T. 2013. "Transplants Keep Rolling in North America." *Automotive News*, April 22.

Leonhart, David. 2008. "$73 an Hour: Adding It Up." *The New York Times*, December 9. www.nytimes.com/2008/12/10/business/economy/10leonhardt.html?_r=4&em&.

Linebaugh, Kate, and Jeff Bennett. 2010. "Marchionne Upends Chrysler's Ways: CEO Decries Detroit's 'Fanatical' Focus on Market Share; Deep Discounts Are Out." *Wall Street Journal*, January 12.

Marchionne, Sergio. 2014. From "Question and Answer Session Two: Inside the Bail-Out," from transcript of the event "Recovery Road? An Assessment of the Auto Bailout and the State of U.S. Manufacturing: A Discussion with Chrysler Chairman and CEO Sergio Marchionne and Larry Summers. Brookings Institution." Event: www.brookings.edu/events/2014/05/21-assessment-of-auto-bailout-us-manufacturing#/full-event/. Transcript: http://www.brookings.edu/~/media/events/2014/5/21%20auto%20bailout/20140521_auto_bailout_transcript.pdf.

Motor & Equipment Manufacturers Association. 2009. Motor Vehicle Supplier Sector Emergency Financial Assistance Request, Submitted to U.S. Department of Treasury, February 13, 2009.

Nardelli, Robert. 2008. Testimony, Committee on Financial Services, U.S. Senate "Stabilizing the Financial Condition of the American Automobile Industry," 110th Congress, Second Session, November 19, 2008. Serial No. 110-146. http://www.gpo.gov/fdsys/pkg/CHRG-110hhrg46594/html/CHRG-110hhrg46594.htm.

Obama, Barack. 2009. "Remarks by the President on General Motors Restructuring." June 1**.** http://www.whitehouse.gov/the_press_office/Remarks-by-the-President-on-General-Motors-Restructuring/.

Rattner, Steven. 2010. *Overhaul: An Insider's Account of the Obama Administration's Emergency Rescue of the Auto Industry*. New York, NY: Houghton Mifflin Harcourt.

Scott, Robert E. 2008. "When Giants Fall: Shutdown of One or More U.S. Automakers Could Eliminate up to 3.3 Million U.S. Jobs." EPI Briefing Paper 227, Economic Policy Institute. December 3.

Shepard, David. 2015. "Obama Heralds U.S. Auto Turnaround." *The Detroit News*, January 7. http://www.detroitnews.com/story/business/autos/2015/01/06/obama-auto-turnaround/67431/.

Stole, John D., and Jeffrey McCracken. 2009. "Bankruptcy Fears Grip Auto-Parts Suppliers." *Wall Street Journal*, January 26.

US Department of the Treasury. 2015. "Auto Industry." Webpage, last updated 1/8/2015: http://www.treasury.gov/initiatives/financial-stability/TARP-Programs/automotive-programs/Pages/default.aspx.

Warren, Elizabeth. 1987. "Bankruptcy Policy." *University of Chicago Law Review* 54: 775

Zandi, Mark. 2008. Testimony before the U.S. Senate Banking Committee, "The State of the Domestic Auto Industry: Part II," December 4.

Journal of Economic Perspectives—Volume 29, Number 2—Spring 2015—Pages 25–52

The Rescue of Fannie Mae and Freddie Mac[†]

W. Scott Frame, Andreas Fuster, Joseph Tracy, and James Vickery

The imposition of federal conservatorships on September 6, 2008, at the Federal National Mortgage Association and the Federal Home Loan Mortgage Corporation—commonly known as Fannie Mae and Freddie Mac—was one of the most dramatic events of the financial crisis. These two government-sponsored enterprises play a central role in the US housing finance system, and at the start of their conservatorships held or guaranteed about $5.2 trillion of home mortgage debt.

Fannie Mae and Freddie Mac are publicly held financial institutions that were created by Acts of Congress to fulfill a public mission: to enhance the liquidity and stability of the US secondary mortgage market and thereby promote access to mortgage credit, particularly among low- and moderate-income households and neighborhoods. Their federal charters provide important competitive advantages that, taken together, implied US taxpayer support of their financial obligations. As profit-maximizing firms, Fannie Mae and Freddie Mac leveraged these advantages over the years to become very large, very profitable, and very politically powerful. The two firms were often cited as shining examples of public-private partnerships—that is, the harnessing of private capital to advance the social goal of expanding homeownership. But in reality, the hybrid structures of Fannie Mae and Freddie Mac were destined to fail at some point, owing to their singular exposure to residential real

■ *W. Scott Frame is Financial Economist and Senior Policy Adviser, Federal Reserve Bank of Atlanta, Atlanta, Georgia. Andreas Fuster is Senior Economist, Joseph Tracy is Executive Vice President and Senior Advisor to the President, and James Vickery is Research Officer, all at the Federal Reserve Bank of New York, New York, New York. Their email addresses are scott.frame@atl.frb.org, andreas.fuster@ny.frb.org, joseph.tracy@ny.frb.org, and james .vickery@ny.frb.org.*

[†] To access the Appendix and Data Appendix, visit http://dx.doi.org/10.1257/jep.29.2.25

doi=10.1257/jep.29.2.25

estate and moral hazard incentives emanating from the implicit guarantee of their liabilities (for a detailed discussion, see Acharya et al. 2011). A purposefully weak regulatory regime was another important feature of the flawed design. While the structural problems with Fannie Mae and Freddie Mac were understood by many, serious reform efforts were often portrayed as attacks on the American Dream of homeownership, and hence politically unpalatable.

In 2008, as the housing crisis intensified, Fannie Mae and Freddie Mac became financially distressed. Their concentrated exposure to US residential mortgages, coupled with their high leverage, turned out to be a recipe for disaster in the face of a large nationwide decline in home prices and the associated spike in mortgage defaults. As financial markets in the summer of 2008 turned against Fannie Mae and Freddie Mac, the federal government initially responded by passing the Housing and Economic Recovery Act (HERA), signed into law on July 30, 2008, which among many other provisions temporarily gave the US Treasury unlimited investment authority in the two firms. Less than two months later, their new regulator, the Federal Housing Finance Agency (FHFA), placed Fannie Mae and Freddie Mac into conservatorship, taking control of the two firms in an effort to curtail the risk of financial contagion and to conserve their value. Concurrently, the Treasury entered into senior preferred stock purchase agreements with each institution. Under these agreements, US taxpayers ultimately injected $187.5 billion into Fannie Mae and Freddie Mac.

This paper begins by describing the business model of Fannie Mae and Freddie Mac and their role in the US housing finance system. Our focus then turns to the sources of financial distress experienced by the two firms and the events that ultimately led the federal government to take dramatic action in an effort to stabilize housing and financial markets. We describe the various resolution options available to US policymakers at the time and evaluate the success of the choice of conservatorship in terms of its effects on financial markets and financial stability, on mortgage supply, and on the financial position of the two firms themselves. Our overall conclusion is that conservatorship achieved its key short-run goals of stabilizing mortgage markets and promoting financial stability during a period of extreme stress. However, conservatorship was intended to be a temporary fix, not a long-term solution. More than six years later, Fannie Mae and Freddie Mac still remain in conservatorship and opinion remains divided on what their ultimate fate should be.

Background

By law, Fannie Mae and Freddie Mac are limited to operating in the secondary "conforming" mortgage market. This terminology means that the two firms can neither lend money to households directly in the primary market, nor deal in mortgages with balances above a certain size—the "conforming loan limits." The conforming loan limits have been adjusted over time, and for 2015 the national limit for single-family properties is $417,000, but can be as high as $625,500 in high-housing-cost areas. Mortgages with principal balances above the conforming loan limits are referred to as "jumbo" loans. Fannie Mae and Freddie Mac are further limited by law to dealing in mortgages with a downpayment of at least 20 percent, or that maintain equivalent credit enhancement via private mortgage insurance or other means. The two firms otherwise define their own underwriting standards in terms of acceptable credit scores, debt-to-income ratios, and documentation.[1]

Fannie Mae and Freddie Mac's activities take two broad forms. First, their "credit guarantee" business involves the creation of residential mortgage-backed securities by purchasing a pool of conforming mortgages from originators—typically banks or mortgage companies—and then issuing a security that receives cash flows from the mortgage pool. For these "agency" mortgage-backed securities, Fannie Mae or Freddie Mac promise investors timely payments of principal and interest, even if there are defaults and losses on the underlying loans. In return for this guarantee, the firms receive a monthly "guarantee fee," effectively an insurance premium coming out of the borrower's interest payment.

Second, the firms' "portfolio investment" business involves holding and financing assets on their own balance sheets, including whole mortgages, their own agency mortgage-backed securities, nonagency mortgage-backed securities, and other types of fixed income securities. Fannie Mae and Freddie Mac largely fund these assets by issuing "agency" debt. The two firms have historically been highly leveraged, with book equity consistently less than 4 percent of total assets. The firms use financial derivatives, such as interest rate swaps, to help manage the market risk associated with their investment portfolios.

Fannie Mae's and Freddie Mac's federal charters provide a range of benefits that result in lower operating and funding costs (see Frame and White 2005 in this journal), such as a line-of-credit with the US Treasury. These advantages, coupled with two past episodes in which the federal government assisted troubled government-sponsored enterprises (US Government Accountability Office 1990, pp. 90–91), served to create a perception in financial markets that agency debt and mortgage-backed securities were implicitly government guaranteed—despite explicit language on these securities stating that they are not US government obligations. As a result, Fannie Mae and Freddie Mac have been able over the decades to issue debt and mortgage-backed securities at lower yields than their stand-alone financial strength ratings would otherwise warrant, by 20 to 40 basis points (Nothaft, Pearce, and Stevanovic 2002; Ambrose and Warga 2002; Passmore 2005).

This funding advantage was partially passed on to borrowers in the form of lower mortgage rates. Econometric studies find that, prior to the financial crisis, conforming mortgages had lower interest rates than jumbo mortgages, with estimates of the gap ranging from 10 to 30 basis points depending on the sample period and estimation approach (for example, Kaufmann 2014; DeFusco and Paciorek 2014; see McKenzie 2002 for a review of earlier literature).

[1] Some mortgages not meeting Fannie Mae or Freddie Mac's underwriting standards may alternatively be financed using government insurance programs (operated by the Federal Housing Administration or Department of Veterans Affairs). Such loans may be securitized with a public credit guarantee to investors via the Government National Mortgage Association (Ginnie Mae) operated by the US Department of Housing and Urban Development.

In 1992, Congress created a two-part regulatory structure to monitor Fannie Mae and Freddie Mac for compliance with their statutory missions and to limit their risk-taking. Mission regulation was assigned to the US Department of Housing and Urban Development (HUD), while safety-and-soundness regulation became the purview of a newly created Office of Federal Housing Enterprise Oversight (OFHEO) as an independent agency within HUD. Congressional placement of OFHEO within HUD can be viewed as a signal that the housing mission goals were the more important priority.

The principal manifestation of mission regulation for Fannie Mae and Freddie Mac was the establishment of affordable housing goals. These goals stipulated minimum percentages of mortgage purchases that finance dwellings in underserved areas and for low- and moderate-income households (see Bhutta 2012 for more details). The goals were progressively increased between 1996 and 2007; for example, the target purchase percentage for low-and-moderate income households was raised from 40 percent to 55 percent during this period. This provided political cover for Fannie Mae and Freddie Mac to expand their business and take on greater risk.

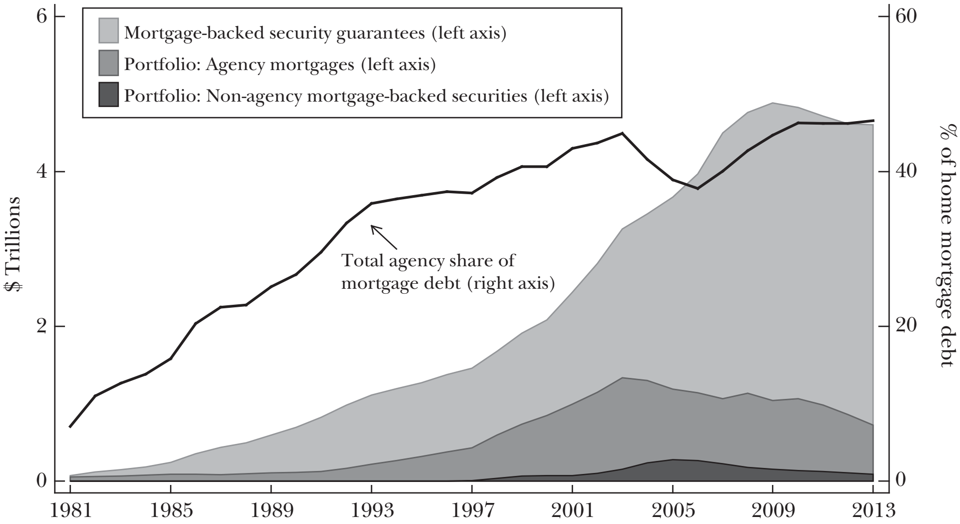
As the safety-and-soundness regulator, OFHEO was authorized to set risk-based capital standards (subject to important statutory limitations), conduct financial examinations, and take certain enforcement actions. However, OFHEO lacked the authority to adjust minimum capital requirements, which were set by statute at very low levels: the sum of 2.5 percent of on-balance sheet assets and 0.45 percent of credit guarantees for agency mortgage-backed securities held by outside investors. The new regulator did not have receivership authority in the event of a failure of either Fannie Mae or Freddie Mac. Finally, OFHEO was subject to the Congressional annual appropriations process and therefore periodically fell victim to political meddling. These and other regulatory deficiencies became clear to many observers (for example, Frame and White 2004 and references therein) but were not addressed until the passage of the Housing and Economic Recovery Act in July 2008.

Figures 1 and 2 highlight the remarkable growth of Fannie Mae and Freddie Mac in recent decades. Figure 1 plots the expansion of the two firms' single-family mortgage credit guarantee and investment portfolios, while Figure 2 plots their cumulative total equity returns compared to the overall market. The stock of agency mortgage-backed securities issued and guaranteed by the two firms (excluding those held by Fannie Mae and Freddie Mac) increased from just $20 billion in 1981 to $3.4 trillion by 2007, the year prior to the start of the conservatorships. Fannie Mae's and Freddie Mac's single-family mortgage investment portfolio holdings (agency mortgages plus nonagency mortgage-backed securities) increased twenty-fold over the same period, from $50 billion to $1.1 trillion. Although the investment portfolios of the two firms have shrunk significantly since they were placed in conservatorship, their total market share inclusive of their mortgage guarantees has continued to grow. The two firms owned or guaranteed 47 percent of single-family mortgage debt outstanding in 2013, compared to 40 percent in 2007 and only 7 percent in 1981. (These figures exclude cross-holdings and ownership of government-guaranteed mortgage assets.)

Fannie Mae and Freddie Mac's share of the mortgage market grew quite steadily between the early 1980s and the early 2000s, although the volume of mortgages they owned or guaranteed accelerated in dollar terms due to overall market growth. The

Figure 1

The Growing Role of Fannie Mae and Freddie Mac in the US Mortgage Market



Sources: US Federal Housing Finance Agency (2014) Annual Report to Congress, Federal Reserve Flow of Funds.

Notes: Figure 1 plots the expansion of the two firms' single-family mortgage credit guarantee and investment portfolios. Statistics reflect single-family mortgages only. The category "Mortgage-backed security guarantees" measures agency mortgage-backed securities held by third parties. To avoid double counting, portfolio holdings exclude cross-holdings (that is, securities issued by either of Fannie Mae or Freddie Mac that are owned by the other). They also exclude government-guaranteed FHA loans. The online Appendix to this paper at http://e-jep.org contains more details about figure construction.

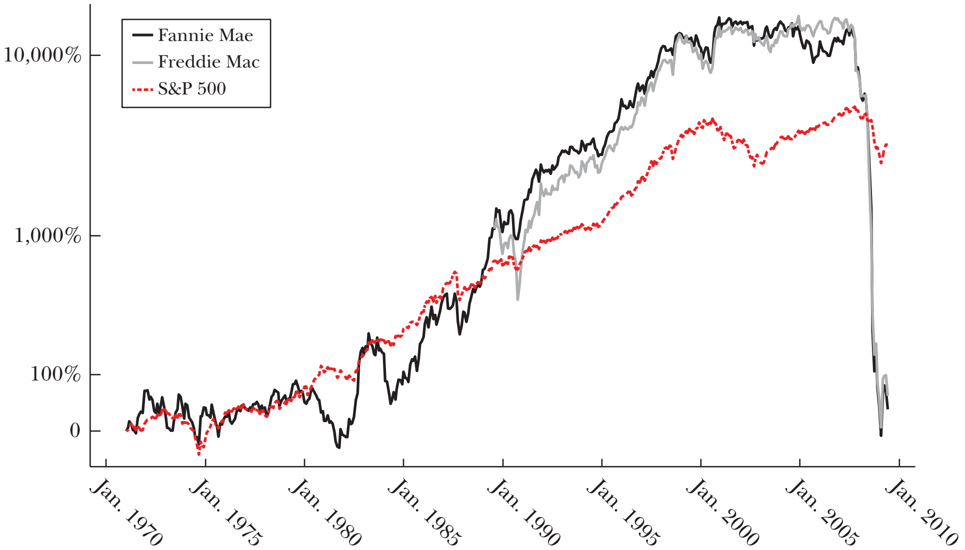
two firms' portfolios of retained mortgage assets, which generate significant additional interest income, grew particularly rapidly from the mid 1990s until the accounting scandals that befell the two firms in 2003 (Freddie Mac) and 2004 (Fannie Mae).

The two firms' growing size and profitability was also reflected in their cumulative stock returns shown in Figure 2. Fannie Mae's stock did not outperform the market in the 1970s and 1980s, and experienced a period of high volatility in the early 1980s due to the high interest rate environment that also triggered the demise of many savings and loan associations (or "thrifts"). (Freddie Mac became publicly traded in 1989.) Both firms significantly outperformed the overall stock market in the 1990s, however. These stock price gains reflected expectations and realizations of rapid, profitable growth, achieved through a combination of mortgage market growth, changes in senior management strategy, a greater understanding of how to leverage their existing funding advantage, and the very low statutory capital requirements established in 1992.[2] The two firms also started competing more directly. Historically,

[2] Demand-side forces likely also played a key role. For example, Basel I risk-based capital regulations gave some banks an incentive to swap their mortgages for agency mortgage-backed securities and encouraged other banks to sell mortgage assets outright. This helped spur the firms' credit guarantee and investment portfolio businesses, respectively (Frame and White 2005).

Figure 2

Cumulative Total Equity Returns of Fannie Mae and Freddie Mac Relative to S&P 500



Source: Center for Research in Security Prices.
Notes: Figure 2 plots the natural logarithm of cumulative returns, inclusive of dividends and other distributions, over the period from January 1971–June 2009. The cumulative return for Freddie Mac is set to be at the same level as Fannie Mae's in August 1989, when our total return series for Freddie Mac starts.

Freddie Mac had securitized mortgages originated by savings and loan institutions, whereas Fannie Mae tended to hold mortgages purchased from mortgage banks, but this segmentation broke down over time.

Fannie Mae and Freddie Mac's stock returns became lower and more volatile after 2002 (recall, the figure shows cumulative returns, so a flat line means essentially zero return). Their accounting scandals resulted in increased capital requirements (so-called capital surcharges) that dampened profitability and triggered legislative reform efforts that created additional uncertainty about the firms' future charter values. The firms also faced greater competition from the rapidly growing nonagency securitization market. Figure 2 also illustrates the rising concerns about financial distress at Fannie Mae and Freddie Mac in 2007 and 2008, and shows how the imposition of the federal conservatorships virtually eliminated the value of common shares of the two firms. We focus on this period in the next section.

While Fannie Mae and Freddie Mac traditionally held or guaranteed prime conforming mortgages with low historical default risk, the activities of the two firms were influenced during the 2000s by the rapid growth in the higher-risk "subprime" mortgage market (for a description of this market, see Ashcraft and Schuermann 2008; Mayer, Pence, and Sherlund 2009, in this journal). Although pools of subprime mortgages were generally turned into securities by investment banks rather than by Fannie Mae and Freddie Mac, the two firms were significant investors in these "nonagency" mortgage-backed securities, which were viewed as profitable investments that also helped satisfy affordable housing goals. By the end of 2007, the two firms owned over $300 billion of nonagency mortgage-backed securities.

There is also some evidence that the riskiness of conforming mortgages owned or guaranteed by Fannie Mae and Freddie Mac increased leading up to 2008, perhaps due to competition from nonagency securitization. For example, at Fannie Mae the percentage of newly purchased loans where the loan amount was 90 percent or more of the appraised property value increased from 7 percent in 2003 to 16 percent by 2007; for Freddie Mac, the corresponding share rose from 5 percent in 2003 to 11 percent in 2007. These statistics likely understate true borrower leverage, due to unreported second loans or "piggyback" mortgages, which became common during the housing boom. The share of loans guaranteed by Fannie Mae and Freddie Mac with nonstandard (and risky) features such as an interest-only period also increased substantially. Subsequent mortgage defaults suffered by the two firms were highly concentrated in the 2005–2008 mortgage vintages.[3]

A range of observers had voiced concerns about the systemic risk posed by Fannie Mae and Freddie Mac some years prior to the financial crisis (for example, Greenspan 2004, 2005), although others suggested the likelihood of an insolvency or liquidity crisis from these firms was very low (for example, Hubbard 2003; Stiglitz, Orszag, and Orszag 2002). The concerns focused on the firms' concentration and hedging of mortgage-related interest rate risk, which seemingly magnified shocks to Treasury and interest rate derivatives markets in the early 2000s (see Eisenbeis, Frame, and Wall 2007 and the references therein).

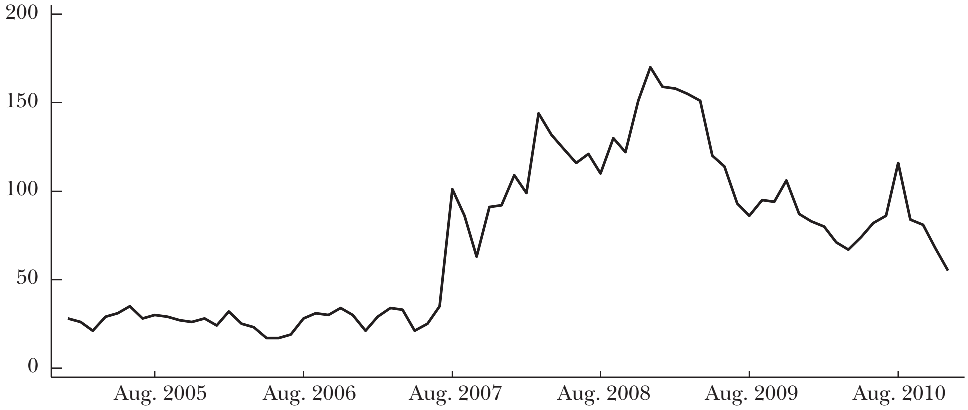
Instead, the two firms were ultimately imperiled by mortgage credit risk, primarily associated with their guarantee activities. The limited attention that policymakers paid to credit risk at Fannie Mae and Freddie Mac was perhaps due to a history of low credit losses on their past guarantees, reflecting both relatively conservative underwriting and a long period of stable or rising home prices. Relatively few observers highlighted the firms' rising exposure to credit risk or anticipated the possibility of a large nationwide decline in home prices.

Events Prior to Conservatorship

US housing and mortgage markets became increasingly stressed during 2007 and 2008 as a result of significant house price declines and the weakening economy. A large number of borrowers found themselves in a situation where the balance on their mortgage exceeded the value of their homes (that is, "negative equity"), which is often a precursor of mortgage default (for example, Foote, Gerardi, and Willen 2008). The tremendous wave of defaults and subsequent foreclosures imperiled many financial institutions with significant exposure to US residential real estate—including Fannie Mae and Freddie Mac. Below, we describe the key events that led

[3] An online Appendix available with this paper at http://e-jep.org, contains statistics about the characteristics of mortgages held or guaranteed by Fannie Mae and Freddie Mac, as well as default rates.

Figure 3
Jumbo–Conforming Spread
(basis points)



Source: Bankrate, Bloomberg Finance L.P.
Notes: Figure 3 shows the unconditional difference in 30-year fixed rate mortgage interest rates between prime jumbo mortgages and conforming mortgages (monthly averages). Jumbo mortgages have a loan amount exceeding the conforming loan limit, making them ineligible for purchase or securitization by Fannie Mae and Freddie Mac.

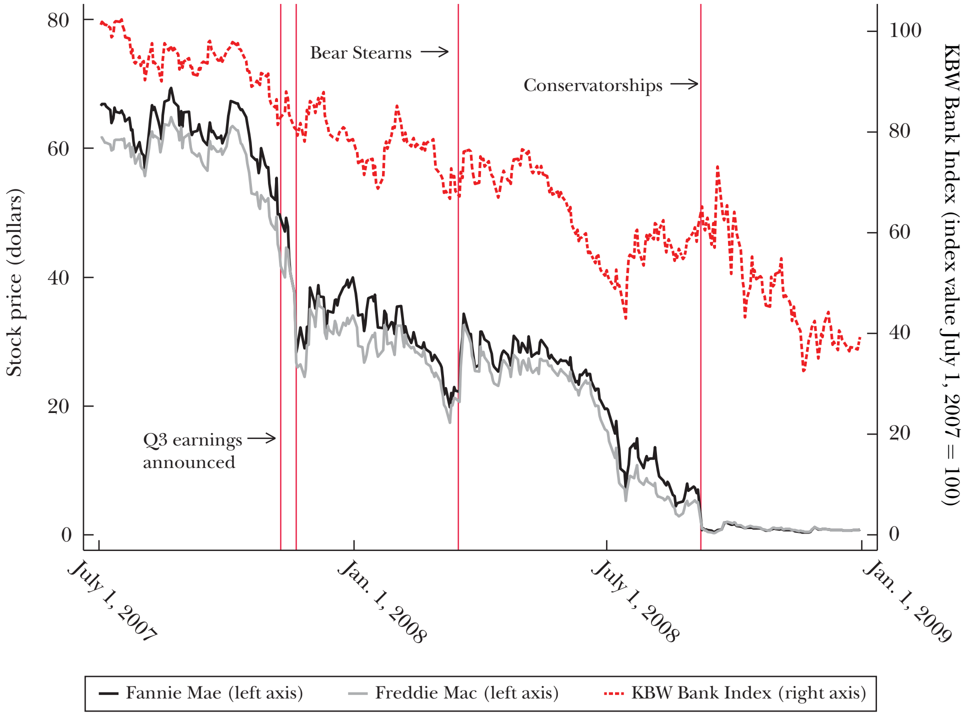
to the conservatorships at these two firms; a detailed chronology is provided in an online Appendix available with this paper at http://e-jep.org.

In summer 2007, as subprime mortgage defaults escalated, issuance of nonagency mortgage-backed securities essentially came to a halt, and other financial markets such as the asset-backed commercial paper market similarly dried up (for discussions of these events, see Brunnermeier 2009, in this journal; Dwyer and Tkac 2009). This period is now widely considered to mark the beginning of the financial crisis. As issuance of nonagency mortgage-backed securities froze, interest rates on prime, but nonconforming, "jumbo" mortgages increased significantly—from about 25 to 100 basis points above those for conforming loans eligible for securitization via the still-liquid agency mortgage-backed securities market, as shown in Figure 3. This historically wide spread between jumbo and conforming mortgages persisted throughout the financial crisis, reflecting both the greater liquidity of conforming mortgages, and the heightened value of the agency credit guarantee. The volume of new jumbo mortgages declined, and the role of Fannie Mae and Freddie Mac expanded as commercial banks became increasingly unwilling or unable to hold new mortgages on their balance sheets (Calem, Covas, and Wu 2013; Fuster and Vickery 2015).

Losses at Fannie Mae and Freddie Mac started mounting: they reported a combined net loss of $8.7 billion during the second half of 2007, reflecting both credit losses on the mortgages they had guaranteed or were holding in portfolio, and mark-to-market losses on their investments. Nevertheless, the two firms' role in the mortgage market further expanded following a temporary increase in conforming loan limits to as high as $729,750 under the Economic Stimulus Act

Figure 4

Fannie Mae and Freddie Mac Stock Prices, July 2007—December 2008



Source: Bloomberg Finance L.P.
Note: Vertical lines mark November 9 and 20, 2007 (when Fannie Mae and Freddie Mac announced their earnings for the 3rd quarter of 2007); March 16, 2008 (Bear Stearns acquisition); and September 7, 2008 (conservatorship announcement).

passed in February 2008 (for details, see Vickery and Wright 2013). Furthermore, during the first quarter of 2008, the Office of Federal Housing Enterprise Oversight removed limits on the size of the investment portfolios at Fannie Mae and Freddie Mac and lowered surcharges to each firm's capital requirements so that they could purchase or guarantee additional mortgages. These portfolio limits and capital surcharges had been imposed by the OFHEO between 2004 and 2006 due to concerns about accounting practices at the two firms.

By mid-2008, after adding over $600 billion in mortgage credit exposure over the previous four quarters, the two firms had expanded to almost $1.8 trillion in combined assets and $3.7 trillion in combined net off-balance sheet credit guarantees. But over the year to June 2008, Fannie Mae and Freddie Mac together posted $14.2 billion in losses and saw their capital recede to $41.2 billion (Fannie Mae) and $12.9 billion (Freddie Mac). At this point, their combined capital amounted to only about 1 percent of their exposure to mortgage risks, a tiny cushion in the face of large expected losses.

Investors became increasingly concerned about the financial condition of Fannie Mae and Freddie Mac during summer 2008. Figure 4 illustrates how their share prices first fell sharply during fall 2007 after both firms reported losses for the third quarter of 2007, and then fell from $25–30 in April 2008 to below $10 in mid-July. Debt investors also increasingly sought clarity from the federal government about whether bondholders would be shielded from losses.

Against this backdrop, and in an effort to calm markets, Treasury Secretary Henry Paulson proposed a plan in July 2008 to allow the Treasury to make unlimited debt and/or equity investments in Fannie Mae and Freddie Mac. (It was in a Senate Banking Committee hearing at this time when Paulson famously stated that "If you've got a bazooka [in your pocket] and people know you've got it, you may not have to take it out" (Paulson 2010).) This plan was incorporated as part of the Housing and Economic Recovery Act, which was signed into law later in July 2008. The law also created the Federal Housing Finance Agency (FHFA), and for the first time granted the new supervisor the authority to place a distressed government-sponsored enterprise into receivership. Immediately following the passage of the new housing legislation, the Treasury began a comprehensive financial review of Fannie Mae and Freddie Mac in conjunction with the FHFA, the Federal Reserve, and Morgan Stanley (Paulson 2010). The Housing and Economic Recovery Act required that FHFA consult with the Treasury and Federal Reserve on any resolution of Fannie Mae or Freddie Mac.

Fannie Mae and Freddie Mac released their second quarter earnings in early August 2008. As shown in Table 1, at this time the two firms were both technically solvent, in the sense that the book value of their equity capital was positive, and indeed exceeded statutory minimum requirements. However, there was a compelling case that, when viewed on an economic basis, both firms were actually insolvent. First, both firms were recognizing large "deferred tax assets" to offset future income taxes ($20.6 billion for Fannie Mae and $18.4 billion for Freddie Mac). Arguably these assets had little immediate value in light of the firms' extremely weak near-term earnings prospects. Excluding these assets, as would have been done for regulatory capital purposes if the two firms had been treated like banks, reduces their measured net worth to $20.6 billion (Fannie Mae) and −$5.5 billion (Freddie Mac). Second, the reported fair market value of their assets (net of liabilities) was significantly lower than book equity, and in Freddie Mac's case was actually negative. Even these fair values may have understated the firms' financial problems, since there is evidence that their accounting reserves against expected future credit losses were also insufficient (US Financial Crisis Inquiry Commission 2011, p. 317). These facts, together with continued deteriorating mortgage market conditions and potential near-term difficulties in rolling over the firms' significant short-term debt (shown in Table 1), created a keen sense of urgency for the US government to take action.

Resolution: Issues, Options, and Actions

Why Was Action Needed?

Our view is that it was appropriate to provide temporary public support for Fannie Mae and Freddie Mac in September 2008. We now present the case

Table 1
Balance Sheet Composition as of June 2008

	Accounting value ($ billions)	
	Fannie Mae	*Freddie Mac*
Assets		
Cash, federal funds, and repurchase agreements	$49.4	$58.8
Investment securities, at fair value	$344.8	$684.7
Agency mortgage-backed securities	$220.4	$490.2
Private-label mortgage-backed securities & revenue bonds	$96.1	$181.6
Other investment securities	$28.3	$12.9
Whole mortgage loans	$418.2	$89.1
Deferred tax assets	$20.6	$18.4
Other assets	$52.9	$28.1
Total assets	**$885.9**	**$879.0**
Liabilities		
Short-term debt (Maturity < 1 year)	$240.2	$326.3
Long-term debt	$550.3	$505.0
Subordinated debt	$9.0	$4.5
Other liabilities	$45.0	$30.2
Total liabilities	**$844.5**	**$866.0**
Equity		
Common stock, other paid-in capital, retained earnings	$32.5	$27.1
Preferred stock	$21.7	$14.1
Treasury stock	($7.3)	($4.1)
Accumulated other comprehensive loss	($5.7)	($24.2)
Total Equity	**$41.2**	**$12.9**
Memo: Off balance sheet credit guarantees (net)	$2,289.9	$1,409.9

Notes: This table provides summarized balance sheet information for Fannie Mae and Freddie Mac as of June 30, 2008. Balance sheet measures are presented at historical cost according to generally accepted accounting principles as reported in each firm's 10-K. Off-balance sheet credit guarantees are from each firm's "monthly summary" and net of their own mortgage-backed securities held on balance sheet. They are contingent liabilities. A more detailed balance sheet is presented in the online Appendix at http://e-jep.org.

for public intervention, drawing on economic theory and information about conditions at the time.

A key argument in favor of intervention was to support the supply of conforming mortgages during a period of severe financial stress. As already discussed, the sharp rise in the spread between jumbo and conforming mortgage interest rates during 2007–2008 was prompted by a freeze in private jumbo securitization, generally attributed to heightened asymmetric information and uncertainty about mortgage credit risk (Leitner 2011). The freeze did not extend to agency mortgage-backed securities because of their implicit government guarantee. Public support of Fannie Mae and Freddie Mac maintained these guarantees and allowed agency securitization to continue and thereby support the supply of conforming mortgages. Theory provides support for the use of public guarantees as a crisis response; as one example, Philippon and Skreta (2012) present a model in which such guarantees are an optimal intervention in markets subject to adverse selection. Securitization was likely particularly important for mortgage supply during this period because of the limited capacity of banks and other financial intermediaries to hold additional mortgages on their balance sheets due to falling capitalization and the failure of several large lenders (see Shleifer and Vishny 1992 for a model studying the effects of limited industry balance sheet capacity).

Was it important to promote mortgage supply during this period given the already high levels of outstanding US mortgage debt? We would argue "yes," for two reasons.

First, mortgage origination was necessary to enable refinancing of existing mortgages. The overall policy response to the financial and economic crisis involved a significant easing of monetary policy, which works in part by lowering interest rates on existing debt contracts. Such a decrease in rates has been found to lower mortgage defaults (Fuster and Willen 2012; Tracy and Wright 2012; Zhu, Janowiak, Ji, Karamon, and McManus forthcoming) and to stimulate consumption (Keys, Piskorski, Seru, and Yao 2014; Di Maggio, Kermani, and Ramcharan 2014). Interest rates on fixed-rate mortgages, which make up the vast bulk of the stock of US mortgage debt, only respond to lower market rates if borrowers can refinance. Even with the rescue of Fannie Mae and Freddie Mac, lower yields on mortgage-backed securities were only partially transmitted to primary mortgage interest rates during this time (Fuster et al. 2013; Scharfstein and Sunderam 2014). But refinancing would almost certainly have been even more difficult without Fannie Mae and Freddie Mac, considering the tight lending standards for nonconforming mortgages at the time.

Second, continued mortgage supply enabled at least some households to make home purchases during a period of extreme weakness in the housing market.[4] A large body of theory models how changes in credit availability can lead to a negative spiral among asset prices, collateral values, and credit availability (for a prominent example, see Kiyotaki and Moore 1997). Consistent with the spirit of such models, Kung (2014) finds empirically that the local increases in the conforming loan limit in 2008, which made more loans eligible for agency securitization, raised home prices by around 6 percent for homes in San Francisco and Los Angeles that were most likely to be purchased with these newly eligible loans.

These arguments support the use of government guarantees in 2008 to help finance new mortgages. But what about the legacy securities issued by Fannie Mae and Freddie Mac prior to September 2008? In our view, if explicit government support of the firms had not been forthcoming, market perceptions of a material credit risk embedded in existing agency debt and mortgage-backed securities could have substantially destabilized the broader financial system given the sheer volume of such securities outstanding, the large holdings of leveraged institutions such as commercial banks, insurance firms, and securities broker-dealers (an online Appendix available with this paper at http://e-jep.org provides statistics about these holdings) and their widespread use as collateral in short-term funding markets.

[4] RealtyTrac (2014) estimates that around 60–65 percent of single-family home purchases in 2009 involved a new mortgage loan, with the remainder going to all-cash buyers.

Credit losses on agency securities would have exacerbated the weak capital and liquidity position of many already-stressed financial institutions and raised the possibility of forced asset sales and runs (as in the models posited by Diamond and Rajan 2011 or Diamond and Dybvig 1983). Finally, Fannie Mae and Freddie Mac held large positions in interest rate derivatives for hedging. A disorderly failure of these firms would have caused serious disruptions for their derivative counterparties.

A further consideration was that almost $1 trillion of agency debt and mortgage-backed securities was held by foreign official institutions, mainly central banks. Allowing these securities to default would likely have had significant international political ramifications.[5] Furthermore, as emphasized by Paulson (2010) and Acharya et al. (2011), given the widespread perception that agency debt and mortgage-backed securities were implicitly government guaranteed, a default by Fannie Mae or Freddie Mac would potentially raise the risk of questions about creditworthiness of the US government, disrupting the US Treasury debt market and increasing the government's funding costs.

Summing up, Fannie Mae and Freddie Mac were too large and interconnected to be allowed to fail, especially in September 2008 given the deteriorating conditions in US housing and financial markets and the central role of these two firms in the mortgage finance infrastructure. Our view is that an optimal intervention would have involved the following elements:

1) Fannie Mae and Freddie Mac would be enabled to continue their core securitization and guarantee functions as going concerns, thereby maintaining conforming mortgage credit supply.
2) The two firms would continue to honor their agency debt and mortgage-backed securities obligations, given the amount and widely held nature of these securities, especially in leveraged financial institutions, and the potential for financial instability in case of default on these obligations.
3) The value of the common and preferred equity in the two firms would be extinguished, reflecting their insolvent financial position.
4) The two firms would be managed in a way that would provide flexibility to take into account macroeconomic objectives, rather than just maximizing the private value of their assets.
5) The structure of the rescue would prompt long-term reform and set in motion the transition to a better system within a reasonable period of time.

Later in the paper, we evaluate actions taken relative to these five objectives, concluding that the path taken was quite successful on the first three, but less successful on the last two.

[5] Paulson (2010, p. 160) discusses learning on his trip to the 2008 Summer Olympics in Beijing that Russian officials had approached the Chinese government about a joint plan to dump a large portion of their holdings of Fannie Mae and Freddie Mac in an effort to create a financial crisis that would force US authorities to support the firms explicitly. For details on these holdings of agency securities, see the online Appendix to this article available with the paper at http://e-jep.org.

What Action Was Taken?

On September 7, 2008, Director of the Federal Housing Finance Agency James Lockhart, Secretary of the Treasury Hank Paulson, and Chairman of the Federal Reserve Ben Bernanke outlined a plan to stabilize the residential mortgage finance market. This included: 1) placing both Fannie Mae and Freddie Mac into conservatorship; 2) having the Treasury enter into senior preferred stock purchase agreements with both firms; and 3) establishing two new Treasury-operated liquidity facilities aimed at supporting the residential mortgage market—a mortgage-backed securities purchase facility and a standing credit facility. We discuss these steps in turn.

By becoming a conservator, the Federal Housing Finance Agency assumed the responsibilities of the directors, officers, and shareholders of both Fannie Mae and Freddie Mac with the purpose of conserving their assets and rehabilitating them into safe-and-sound condition. Hence the two institutions would continue as going concerns, carry out their usual market functions, and continue to pay their financial obligations. The boards of Fannie Mae and Freddie Mac consented to the appointment of the conservator, although the chief executive officers and directors of each firm were then immediately replaced.

The US Treasury's senior preferred stock purchase agreements sought to ensure that Fannie Mae and Freddie Mac maintained positive net worth going forward. Under the agreements, if the Federal Housing Finance Agency determines that either institution's liabilities exceed their assets under generally accepted accounting principles (GAAP), the Treasury would contribute cash capital equal to the difference, in exchange for senior preferred stock. (Specifically, this preferred stock is senior to the prior existing common and preferred equity of the two firms, but junior to their senior and subordinated debt and mortgage-backed securities.) Each agreement was initially for an indefinite term and for up to $100 billion, although the maximum was raised by subsequent amendments to $200 billion per enterprise in February 2009, then in December 2009 to an unlimited amount through the year 2012. As we discuss in more detail later, under these agreements the two firms jointly ended up drawing a total of $187.5 billion over the course of 2008 to 2011.

The senior preferred stock accrued dividends at 10 percent per year. The senior preferred stock purchase agreements also required both Fannie Mae and Freddie Mac to provide the Treasury with: 1) $1 billion of senior preferred shares; 2) warrants that would allow the purchase of common stock representing 79.9 percent of each institution on a fully diluted basis;[6] and 3) a quarterly commitment fee to be determined by the Treasury and the Federal Housing Finance Agency (as conservator) in consultation with the Federal Reserve.[7] To date, the Treasury has not exercised the warrants

[6] The 79.9 percent ownership stake was selected to avoid the necessity to consolidate the assets and liabilities of Fannie Mae and Freddie Mac onto the government's balance sheet. See Swagel (2009, p. 37).

[7] The senior preferred stock purchase agreements also included various covenants. Specifically, Treasury approval is required before: 1) purchasing, redeeming or issuing any capital stock or paying dividends;

to purchase common stock. In accordance with the terms of the agreement, Treasury waived the commitment fee each period, and then suspended this provision in 2012.

The senior preferred stock purchase agreements also required Fannie Mae and Freddie Mac to begin winding down their retained investment portfolios, starting in 2010, at a rate of at least 10 percent per year until they each fall below $250 billion. This provision was intended to assuage policymaker concerns that these investment portfolios might pose future systemic risk to the financial system.

In September 2008, the US Treasury also created a Government Sponsored Enterprise Credit Facility in which Fannie Mae, Freddie Mac, and the Federal Home Loan Bank System could borrow on a short-term collateralized basis from the Treasury. The facility was never used and expired on December 31, 2009. The Treasury furthermore introduced a temporary Mortgage-Backed Securities Purchase Program under which it could purchase agency mortgage-backed securities in an effort to support the mortgage market. It ultimately acquired $225 billion of these securities, which were subsequently sold in 2011 and 2012.

In August 2012, an amendment to the senior preferred stock purchase agreement was announced, in which the fixed 10 percent dividend on the senior preferred stock owned by Treasury was replaced with a "full income sweep." This implied that all profits made by the two firms would be remitted to Treasury, preventing them from building up positive capital (except for a small net worth "buffer" capped at $3 billion per firm and declining over time). Furthermore, the amendment accelerated the reduction of their investment portfolios, going from a wind-down rate of 10 percent per year to 15 percent. When announcing the amendment, the US Department of Treasury (2012) was explicit that a main goal was to "expedite the wind down of Fannie Mae and Freddie Mac."

Why Conservatorship? What Were the Alternatives?

As "federal instrumentalities," Fannie Mae and Freddie Mac are exempt from the bankruptcy code. However, since its creation in 1992, the Office of Federal Housing Enterprise Oversight had the authority to place Fannie Mae or Freddie Mac into "conservatorship" in an effort to conserve their assets and restore them to a safe-and-sound financial condition. The 1992 law, though, did not provide OFHEO either with any funding to assist with a conservatorship, or with a mechanism to fully resolve financial distress at either firm by apportioning losses to shareholders and creditors (Wall, Eisenbeis, and Frame 2005). Under these constraints, a conservatorship ends up looking a lot like "regulatory forbearance"—that is, allowing distressed firms to violate regulations in order to maintain their operations in the hope that they will grow back to financial health.

The Housing and Economic Recovery Act enacted in July 2008 expanded the supervisory options available. First, the law granted receivership authority to the

2) terminating conservatorship other than in connection with receivership; 3) increasing debt to greater than 110 percent of that outstanding as of June 30, 2008; or 4) acquiring, consolidating, or merging into another entity.

newly created Federal Housing Finance Agency.[8] This authority extends those of a conservator by allowing the supervisor to liquidate assets and/or restructure the firm in an effort to limit taxpayer losses. However, formally extinguishing the firms would require Congress to revoke their charters. Absent Congressional action, receivership for either firm would require the creation of a limited life entity (a "bridge entity" akin to a "bridge bank" used when the Federal Deposit Insurance Corporation puts a bank into receivership) that would be financially viable and could maintain the Congressional charter.[9]

Second, as mentioned above, the Housing and Economic Recovery Act of 2008 provided the US Treasury with authority to make unlimited investments in securities of Fannie Mae and Freddie Mac conditional on an "emergency determination" by the Treasury Secretary and agreement from the firm(s) on the terms and conditions of the investment. This investment authority was provided temporarily, through the end of 2009.

Once the federal government decided to rescue Fannie Mae and Freddie Mac and to invest public money, the choice was whether to utilize receivership or conservatorship. This choice became principally about which classes of creditors or shareholders would be made to suffer losses. (For the reasons outlined at the beginning of this section, it seemed unwise in the middle of a financial crisis to follow a course of action that would impose losses on holders of agency debt or mortgage-backed securities.) In the case of conservatorship, US Treasury purchases of common equity would restore the two firms to financial health but would represent a public bail-out of all claimants. Alternatively, the Treasury could purchase a more senior class of securities, which would benefit holders of even more senior obligations but largely wipe out the value of junior obligations. With a receivership, government funding could be used to capitalize the "bridge" entity in an effort to support senior creditors and any other claimants that the government wanted to protect. Subsequently, the Treasury would be expected to hold an initial public offering for the bridge entity in an effort to monetize the taxpayers' investment. Indeed, the Housing and Economic Recovery Act required that the bridge entity

[8] The idea of providing the supervisor of Fannie Mae and Freddie Mac with receivership authority had been debated in the years prior to the financial crisis. Some policymakers, including those at the Federal Reserve and Treasury Department, viewed this as a way to impose greater market discipline on Fannie Mae and Freddie Mac by exposing their bondholders to potential loss. Of course, this increased market discipline would be conditional on receivership being viewed as a credible alternative by the markets. Many legislators, however, were concerned that such supervisory authority would raise the cost of housing finance.

[9] In the absence of any government funding, a receivership utilizing a "bridge" structure would generally work in the following way. The Federal Housing Finance Agency would first evaluate the current and expected performance of the assets and off-balance sheet credit guarantees. "Good assets" expected to perform would then be transferred to the new bridge entity, with the "bad assets" remaining with the original institution. The difference in value between the good and bad assets plus the amount of required capital would represent the amount of loss to be apportioned to claimants in order of priority within the original capital structure: that is, common stockholders, preferred stockholders, subordinated bondholders, and senior bondholders. Mortgage-backed securities investors would maintain their interest in the underlying loans with any shortfall treated as a senior unsecured claim.

be sold within two years of creation (although it includes an option to extend this period by up to three years).

If the US Treasury had not received financing authority in the Housing and Economic Recovery Act, receivership would likely have provided the better opportunity for ultimately stabilizing the mortgage market. However, given the depth of the problems at Fannie Mae and Freddie Mac, receivership would likely have involved some losses being borne by senior creditors (that is, holders of agency debt and mortgage-backed securities) and a breach of the implicit government guarantee. Conditional on Treasury financing, there were several reasons why the conservatorship was preferable to receivership.

First, in the summer of 2008, there was significant uncertainty about the housing market and future losses at Fannie Mae and Freddie Mac. The presence of this uncertainty meant that, given the time frame allowed, restructuring the two firms via receivership would entail some risk that they could potentially fail again. Hence, receivership might not have solved the critical near-term problem.

Second, the business model of the government-sponsored enterprises had been the subject of intense debate in the years leading up to their failure. The structure of the conservatorship agreements essentially placed Fannie Mae and Freddie Mac in a "time-out." Receivership, by contrast, would have reorganized and released the two firms (at least within five years). The thinking at the time was that conservatorship would force Congress to address the problems of this business model, or else face the long-term prospect of government control of the US housing finance system.

Third, receivership raised an operational concern relating to the treatment of derivatives as "qualified financial contracts" (as discussed by Paulson 2010). Receivership required a determination within one business day about the status of individual counterparties: specifically, whether their claims would be transferred to the "good" entity or remain with the "bad" entity. Depending on that determination, counterparties held the option to terminate net positions. Under law, however, the conservatorship did not trigger these termination options in derivatives contracts (US Federal Housing Finance Agency 2008). Thus, receivership would have created greater uncertainty about business continuity and derivatives counterparty actions.

Finally, conservatorship still allowed for the receivership option to be chosen in the future if a subsequent administration felt that it was a better course of action.

Another alternative option was to nationalize Fannie Mae and Freddie Mac, by buying more than 80 percent of the firms' equity and thereby taking a controlling interest. However, as Paulson (2010) describes in his book, the Bush administration was opposed to nationalization or anything that looked like open-ended government involvement. Relative to conservatorship, nationalization would have given the administration more direct control over Fannie Mae and Freddie Mac but would have required the firms to be put on the government's balance sheet. The 2012 "full income sweep" amendment discussed above effectively narrows the difference between conservatorship and nationalization by transferring essentially all profits and losses from the firms to the Treasury.

Could the US Treasury, instead of taking control of (or liquidating) Fannie Mae and Freddie Mac, have calmed financial markets by simply buying up large quantities of agency debt and mortgage-backed securities? Direct purchases could have removed material risk from the financial institution balance sheets. However, a resolution of the financial distress at Fannie Mae and Freddie Mac would still have been necessary in order to ensure continued mortgage credit availability. The sheer quantity of agency securities outstanding, around $5 trillion in total, would also have made a repurchase program challenging or impossible to implement in practice, given the limited time frame. Such a program would have needed to be much larger than the Troubled Asset Relief Program later used to recapitalize banks.

Effects of the Conservatorship

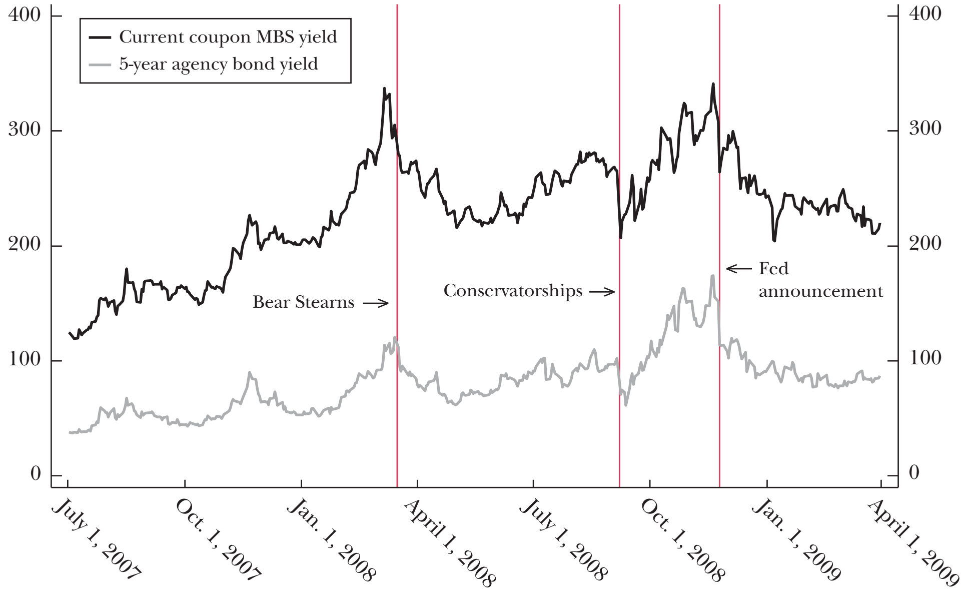
Effects on Financial Markets

The intent of the senior preferred stock purchase agreements and Treasury liquidity facilities was to maintain the firms' operations and to provide assurances to holders of Fannie Mae's and Freddie Mac's debt and mortgage-backed securities. By extension, these actions were expected to both lower and stabilize the cost of mortgage finance. Figure 5 illustrates the announcement effect of the actions taken by looking at the yields of Fannie Mae five-year debt and "current coupon" mortgage-backed securities, both in terms of spreads to five-year Treasury bonds. On the first trading day following the conservatorship announcement, these spreads fell by about 30 basis points (five-year debt) and 50 basis points (mortgage-backed securities). In turn, the fall in mortgage-backed securities yields was followed by a decline in conforming mortgage rates by about 40 basis points within one week. Thus, in the months prior to the announcement, the risk of a potential default by Fannie Mae and Freddie Mac seems to have substantially increased their funding costs and the cost of mortgage credit. At least in the short run, the conservatorship announcement calmed the fears of investors.

As would be expected, the agreements through which the government received preferred stock had significant negative consequences for the existing stockholders. Fannie Mae and Freddie Mac common shares quickly fell below $1 (down from $60 just 12 months earlier), and the Federal Housing Finance Agency subsequently directed both firms to delist from the New York Stock Exchange. Preferred shares suffered a similar fate. Indeed, several community banks became financially distressed as a result of having to write-down the value of their holdings of preferred stock in the two firms (Rice and Rose 2012). Perhaps surprisingly, the two firms maintained their payments on the relatively small amount of subordinated debt that they had outstanding.

The positive bond market reaction, coupled with a relatively smooth operational transition, suggested that the conservatorships at Fannie Mae and Freddie Mac were a success, at least initially. However, as the financial crisis intensified later in the fall of 2008 in the wake of the Lehman Brothers bankruptcy and other events,

Figure 5
Yields on Fannie Mae Debt and Mortgage-Backed Securities (MBS), July 2007–March 2009
(spread in basis points relative to five-year Treasury bonds)

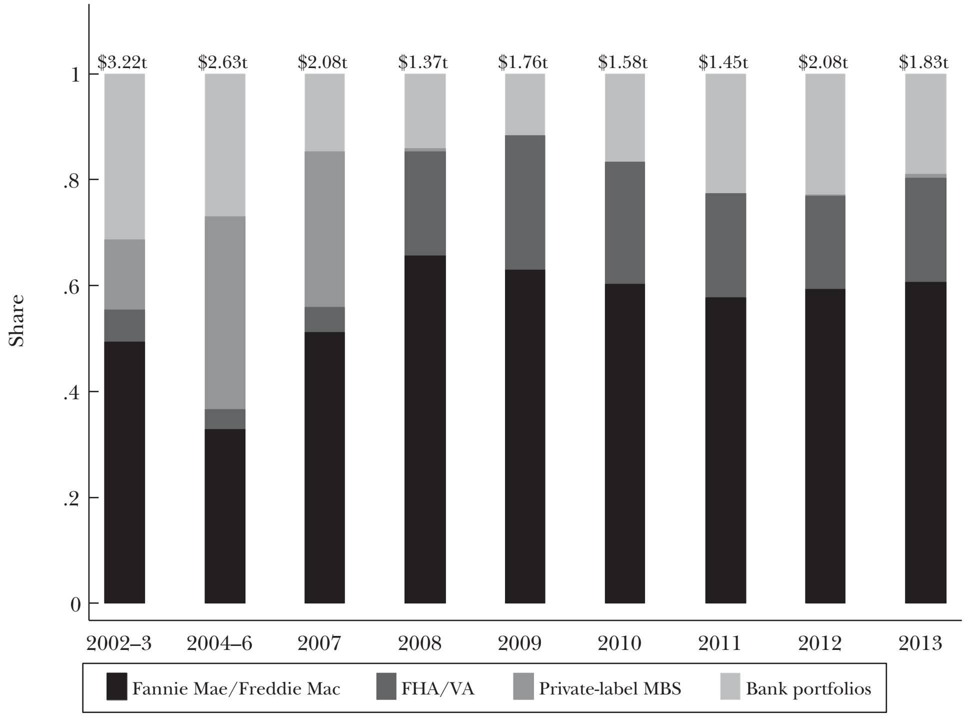


Sources: J.P. Morgan Chase, FRED (Federal Reserve Bank of St. Louis).
Notes: Figure 5 shows the yields of Fannie Mae five-year debt and "current coupon" mortgage-backed securities, both in terms of spreads to five-year Treasury bonds. Vertical lines mark March 16, 2008 (Bear Stearns acquisition); September 7, 2008 (conservatorship announcement); and November 25, 2008 (Fed asset purchase announcement). "Current Coupon MBS" refers to yield of hypothetical mortgage-backed security (MBS) trading at par (see Fuster et al., 2013, for details). The gap between MBS yields and Treasury or swap yields after accounting for the value of the embedded prepayment option (the "option-adjusted spread") displayed qualitatively similar patterns over this period (not shown).

yields on Fannie Mae and Freddie Mac obligations climbed back and soon exceeded their pre-conservatorship levels. This increase appears to have resulted primarily from a general flight to liquidity as well as tight financing conditions during the fall of 2008, rather than a reassessment by the market of what conservatorship would imply for the credit risk of the two firms' bonds going forward (as Krishnamurthy 2010 explained in this journal).

Regardless of the cause, the attendant increase in mortgage rates worried policymakers and became an important contributor to the Federal Reserve's decision to engage in a "large-scale asset purchase program"—commonly referred to as "quantitative easing." On November 25, 2008, the Fed announced that it would purchase up to $500 billion of agency mortgage-backed securities and up to $100 billion of agency debt. As shown in Figure 5, this announcement substantially reduced yield spreads for agency securities, which subsequently normalized over the first quarter of 2009. (For discussions of the channels through which the large-scale

Figure 6
Shares of Different Funding Channels for Newly Originated Mortgages



Source: Inside Mortgage Finance.
Notes: Numbers at the top of each bar indicate total first-lien issuance for the year in trillions of dollars (in case of 2002–2003 and 2004–2006, these are annual averages). "FHA/VA" stands for Federal Housing Administration and the Veterans Administration, which are government agencies that insure loans that are then securitized in Ginnie Mae mortgage-backed securities. "MBS" stands for mortgage-backed securities.

asset purchases affected financial markets, see Gagnon, Raskin, Remache, and Sack 2011; Hancock and Passmore 2011; or Krishnamurthy and Vissing-Jorgensen 2011.) Even though the Fed intervention appears to have lowered yield spreads, this does not mean that, had it come earlier, such an intervention would have stabilized Fannie Mae and Freddie Mac, as the underlying solvency issue would not have been addressed. Indeed, it seems likely that restoring the financial condition of Fannie Mae and Freddie Mac was an important precondition for the Federal Reserve to have been willing to purchase agency securities in the first place.

Effects on Mortgage Lending

Following the decrease in conforming mortgage rates in late 2008, mortgage originations (primarily refinancings) surged, as did issuance of agency mortgage-backed securities, since the conservatorship enabled the credit guarantee businesses of Fannie Mae and Freddie Mac to continue uninterrupted. As shown in Figure 6, since 2008, Fannie Mae and Freddie Mac have guaranteed around

60 percent of originated mortgages, the Federal Housing Administration and the Veterans Administration have insured about 20 percent (securitized by Ginnie Mae), with the remainder held as whole loans by commercial banks. Private-label residential mortgage securitization, which funded more than one-third of mortgages over 2004–2006, has remained close to zero since 2008. Fannie Mae and Freddie Mac's market share is thus higher than ever and almost twice what it was during the height of the housing boom.

The credit profile for Fannie Mae and Freddie Mac's new business has improved since the crisis, as illustrated by the fact that the average credit score on newly guaranteed single-family mortgages increased from below 720 in 2006–2007 to around 760 since 2009 on a scale from 300 to 850 (US Federal Housing Finance Agency 2013). An important reason for this increase in credit scores is that Fannie Mae and Freddie Mac in early 2008 introduced "loan level price adjustments," which are risk-based up-front fees determined by the loan-to-value ratio and the borrower's credit score. These up-front fees have contributed to a steady increase in the overall guarantee fees for new mortgages. For example, Fannie Mae's average effective guarantee fee on new loans tripled from 21 basis points in the first quarter of 2009 to 63 basis points in the first quarter of 2014. Of this increase, 10 basis points was mandated by Congress to fund the 2012 payroll tax reduction.

The Composition of Losses and the Return to Profitability

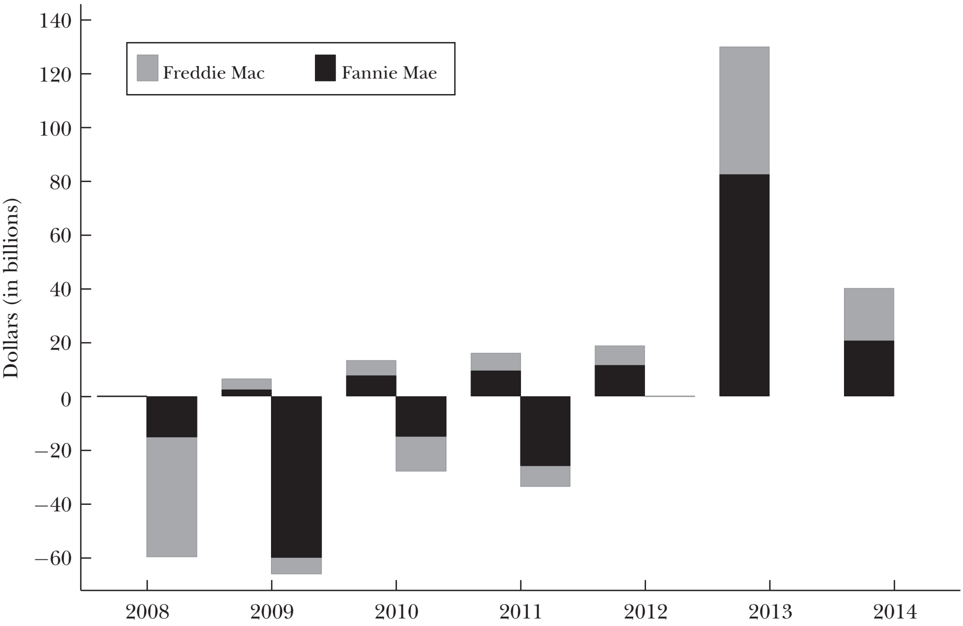
Figure 7 shows the financial consequences of the rescue for the US Treasury. The negative bars show the annual draws by Fannie Mae and Freddie Mac under the senior preferred stock purchase agreements, while the positive bars show the dividends paid. Over the first years of the conservatorship, both firms required very substantial support, but more recently, they have remitted large dividend payments back to the US Treasury.

From 2008 to 2011, Fannie Mae and Freddie Mac posted total combined losses (in terms of comprehensive income) of $266 billion and required $187.5 billion of Treasury support. The biggest contributor to these staggering losses was single-family credit guarantees, which generated about $215 billion in losses over this period, almost all due to provisions for credit losses (US Federal Housing Finance Agency 2011).[10] A second contributor was the dividends on the senior preferred stock held by the US Treasury (paying 10 percent per year), which totaled $36 billion over this period. Perhaps surprisingly, Fannie Mae's and Freddie Mac's investment portfolios, which at first had suffered large losses ($83 billion in 2008), actually generated $2 billion in comprehensive income over this entire period.

In 2012, as house prices stabilized and delinquency rates declined, both Fannie Mae and Freddie Mac stopped losing money on their credit guarantees. Given that their investment portfolios were again profitable, the firms together earned $16 billion

[10] Single-family credit guarantees reflect both guarantees of the firms' agency mortgage-backed securities and whole loans retained on their balance sheets. While losses on the former exceeded the latter, exactly quantifying the two is difficult due to a change in accounting rules in 2010 (US Federal Housing Finance Agency, Office of Inspector General 2012).

Figure 7
Annual Treasury Draws and Dividend Payments, 2008–2014



Source: Fannie Mae and Freddie Mac Financial Results Releases, 3rd quarter of 2014.
Notes: Negative numbers represent draws by Fannie Mae and Freddie Mac, positive numbers represent dividends paid to Treasury. Draws and dividend payments occur one quarter after profits or losses are made.

(after dividend payments to the Treasury). This money was subsequently remitted to the Treasury under the full income sweep amendment to the senior preferred stock purchase agreements noted earlier, which became effective in January 2013.

One consequence of the firms' return to profitability was that their deferred tax assets (which are used to offset taxable income) became useable, and were revalued. As a result, Fannie Mae posted a record profit of $58.7 billion in the first quarter of 2013, and the same happened for Freddie Mac in the third quarter ($30.4 billion). The firms jointly paid dividends of $130 billion to the Treasury during 2013. As of end-2014, the cumulative Treasury dividend payments by Fannie Mae and Freddie Mac have now exceeded their draws: specifically, Fannie Mae has paid $134.5 billion in dividends in comparison to $116.1 billion in draws, while Freddie Mac has paid $91.0 billion in dividends in comparison to $71.3 billion in draws.

Should these figures be interpreted to mean that the Treasury, and therefore taxpayers, have been "repaid" by Fannie Mae and Freddie Mac, and that the two firms should now pay dividends to their regular shareholders again? The answer is no. As an economic matter, one cannot simply compare nominal cash flows but must also take into account that the Treasury took on enormous risk when rescuing the two firms in 2008 and should therefore earn a substantial risk premium, similar to what private investors would have required at the time, in addition to the regular required return (Wall 2014). Furthermore, the effective guarantee has lowered funding costs for Fannie Mae and Freddie Mac and thereby directly contributed to their profits. The US Congressional Budget Office (2010) took these factors into consideration when calculating the total subsidy provided to the firms. Finally, as indicated earlier, the Treasury never collected its commitment fee, which if fairly priced and paid would have significantly reduced the earnings of the two firms. That said, there is some controversy surrounding these issues. In particular, several shareholder lawsuits are contesting the legality of the "sweep" amendment, although with little success to date.[11]

Evaluating the Conservatorships

Earlier, we outlined five desirable objectives of an optimal intervention in response to Fannie Mae and Freddie Mac's financial distress. We believe that the conservatorships largely accomplished the first three objectives, relating to short-run financial stability and credit supply. First, the conservatorships, and particularly the financial support provided by the US Treasury, enabled Fannie Mae and Freddie Mac to support mortgage supply through the crisis and its aftermath. Second, holders of agency debt and mortgage-backed securities did not suffer credit losses (despite the substantial defaults by individual mortgage borrowers), insulating the broader financial system from contagion effects due to the failure of the two firms. Third, both common and preferred equity holders were effectively wiped out, consistent with market discipline. Inconsistent with this objective, however, subordinated debt did not experience losses. While this debt represented only a small part of the liability structure of the two firms, allowing subordinated debt holders to suffer losses may have been desirable in signaling that such debt is indeed risky, thereby curbing moral hazard in similar institutions going forward.

The conservatorship structure was arguably less successful on the fourth objective of aligning the activities of Fannie Mae and Freddie Mac with broader macroeconomic objectives during the Great Recession. The key mission of the conservatorships is to return the two firms to financial health. One year into the conservatorships, Federal Housing Finance Agency Director Lockhart (2009) noted: "We recognize that FHFA's duties as conservator means just that, conserving the Enterprises' assets. This is our top goal."

This focus on the financial performance of the two firms conflicted to some degree, however, with other public policy objectives during this period. One example of this ongoing tension is that, following conservatorship, Fannie Mae and

[11] At the time of this writing, the most recent relevant judgment was that on September 30, 2014: Judge Royce Lamberth of the US District Court for the District of Columbia dismissed several of these claims, based on the view that the Housing and Economic Recovery Act of 2008 empowered Treasury and the Federal Housing Finance Agency to change the terms of the senior preferred stock agreements in this manner. Lamberth's Memorandum Order is at https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2013mc1288-46.

Freddie Mac aggressively enforced "representations and warranties" made by entities that had sold mortgages to them. In practice, the two firms tried to "put back" defaulted mortgages to the originator or seller of the loan, forcing that entity to bear the credit losses.[12] This action was typically justified by flaws in the original documentation or loan underwriting, although importantly, it is not required that the defect be shown to have contributed to mortgage defaults. A consequence of this approach is that the fear of violating representations and warranties on new loans has been cited (especially by originators) as a contributing factor behind tight underwriting standards and higher costs of mortgage lending since the financial crisis (Goodman and Zhu 2013). This tightening of mortgage credit supply has not been helpful to the ongoing recovery of the housing market.

A second example is the role of "principal writedown" (a certain percentage of the borrower's mortgage balance is forgiven) as a policy tool. By the fourth quarter of 2009, an estimated 11.3 million mortgages or 24 percent of borrowers were in negative equity (First American CoreLogic 2010). Borrowers with negative equity are more likely to default, and to produce larger default losses. Such defaults can generate negative externalities, such as reducing prices of nearby properties (Campbell, Giglio, and Pathak 2011). In addition, many argued that the larger issue of debt overhang contributed to lower consumption and created a persistent headwind to economic growth (for example, Mian and Sufi 2014). Absent an explicit policy to address mortgage-related negative equity, this debt overhang would only unwind slowly over time through foreclosures, debt amortization, and any future home price appreciation.

The primary federal program for assisting mortgage borrowers at risk of default was the Home Affordable Modification Program (HAMP), introduced in 2008. Initially, HAMP focused on reducing mortgage payments through reducing interest rates and extending loan terms. Some argued, however, that principal writedown could be a more effective intervention for underwater borrowers (Haughwout, Okah, and Tracy 2010; for an alternative view, see Adelino, Gerardi, and Willen 2014; Eberly and Krishnamurthy 2014). In June 2010, the Treasury expanded HAMP to include a "principal writedown alternative," known as HAMP-PRA. The Federal Housing Finance Agency decided that Fannie Mae and Freddie Mac would not participate in this program, however, due to moral hazard concerns (Fannie Mae 2012). Putting aside the relative merits of principal writedown as a policy tool, what is instructive is the contrast between the broader housing policy perspective of the Treasury versus the FHFA's narrower financial performance goals. In his book, former Treasury Secretary Geithner (2014) recalls: "It was amazing how little actual authority we had over Fannie and Freddie, considering they were entirely dependent on Treasury's cash to stay alive."

[12] Fannie Mae estimates that 3.7 percent of single-family loans acquired between 2005 and 2008 were put back to lenders (source: Fannie Mae 10-K 2013, p. 143). The Federal Housing Finance Agency has also reached a number of settlements with financial institutions related to securities law violation or fraud involving private-label securities purchased by Fannie Mae and Freddie Mac during the boom, totaling more than $16 billion as of mid-2014 (http://www.fhfa.gov/Media/PublicAffairs/Pages/FHFAs-Update-on-Private-Label-Securities-Actions.aspx).

The conservatorships to date have also strikingly failed in relation to our fifth and final objective of producing long-term mortgage finance reform. As Paulson (2010) writes in his book, "We described conservatorship as essentially a 'time out,' or a temporary holding period, while the government decided how to restructure the [government-sponsored enterprises]." However, starting the conservatorships turned out to be easier than ending them, and the "time out" has now stretched into its seventh year.

On February 11, 2011, the US Treasury and Department of Housing and Urban Development (2011) issued a joint white paper on residential mortgage reform. In a press release, Treasury Secretary Geithner described the white paper as follows: "This is a plan for fundamental reform to wind down the [government-sponsored enterprises], strengthen consumer protection, and preserve access to affordable housing for people who need it." But the white paper was only a plan to develop a plan. While the paper outlined three broad possible alternatives for reform, it offered only options without specifics.

Although there appears to be broad consensus that Fannie Mae and Freddie Mac should be replaced by a private system—perhaps augmented by public reinsurance against extreme tail outcomes—substantial disagreement remains about how to implement such a system. The many legislative proposals to date all reflect the cross-currents of trying to protect the taxpayer, preserve support for the 30-year fixed rate mortgage, and keep homeownership affordable to a wide spectrum of borrowers.[13] As yet, there is still no agreed-upon plan for the future of residential mortgage finance.

Conclusions and the Road Ahead

The public actions taken to support Fannie Mae and Freddie Mac were successful in their short-term aims of supporting the housing market and removing the two firms as an immediate source of systemic risk to the financial system. However, the conservatorships have not yet achieved the goal of reforming the system of residential mortgage finance.

The path forward for reform of Fannie Mae and Freddie Mac does not look promising. As time passes since September 2008, the perceived urgency for reform seems to recede. Delay prolongs the uncertainty over the government's future role in residential mortgage finance, which in turn is a deterrent to private capital re-entering the market, and makes the government's role appear more difficult to replace. Delay also raises the likelihood that deeper reform will be judged as too difficult to accomplish, and raises the risk that the conservatorships are ended by returning Fannie Mae and Freddie Mac to private status with only minor changes to

[13] In the US Senate in 2014, the Housing Finance Reform Act of 2013 (S.1217) sponsored by then-Banking Committee Chairman Tim Johnson (D-SD) and Ranking Member Mike Crapo (R-ID) passed through the Banking Committee. However, it is unclear whether this bill can provide the framework for a future reform bill. The current Banking Committee Chairman Senator Richard Shelby (R-AL) voted against the bill, and it is unclear how much support the bill would find in the House of Representatives.

their charters. That is, the key recommendation of the US Treasury and US Department of Housing and Urban Development (2011) white paper—that Fannie Mae and Freddie Mac should be wound down—would in fact not come to pass. This outcome would be a colossal missed opportunity to put US residential mortgage finance on a more stable long-term footing.

■ *The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Atlanta, the Federal Reserve Bank of New York, or the Federal Reserve System. We thank our discussant Amir Sufi and participants at the symposium held by the Journal of Economic Perspectives at the University of Chicago for their thoughtful comments. We also received helpful suggestions from many others, including Adam Ashcraft, David Autor, Mike Fratantoni, Kris Gerardi, Laurie Goodman, Joseph Gyourko, Chang-Tai Hsieh, Wayne Passmore, David Scharfstein, Timothy Taylor, Larry Wall, Larry White, Paul Willen, and Joshua Wright. We also thank Karen Shen and Ulysses Velasquez for research assistance. Any remaining errors are the authors' responsibility.*

References

Acharya, Viral V., Matthew Richardson, Stijn van Nieuwerburgh, and Lawrence J. White. 2011. *Guaranteed to Fail: Fannie Mae, Freddie Mac and the Debacle of Mortgage Finance.* New Jersey: Princeton University Press.

Adelino, Manuel, Kristopher Gerardi, and Paul S. Willen. 2014. "Why Don't Lenders Renegotiate More Home Mortgages? Redefaults, Self-Cures, and Securitizations." *Journal of Monetary Economics* 60(7): 835–53.

Ambrose, Brent W., and Arthur Warga. 2002. "Measuring Potential GSE Funding Advantages." *Journal of Real Estate Finance and Economics* 25(2–3): 129–50.

Ashcraft, Adam B., and Til Schuermann. 2008. "Understanding the Securitization of Subprime Mortgage Credit." *Foundations and Trends in Finance* 2(3): 191–309.

Bhutta, Neil. 2012. "GSE Activity and Mortgage Supply in Lower-Income and Minority Neighborhoods: The Effect of the Affordable Housing Goals." *Journal of Real Estate Finance and Economics* 45(1): 238–61.

Brunnermeier, Markus K. 2009. "Deciphering the Liquidity and Credit Crunch 2007–2008." *Journal of Economic Perspectives* 23(1): 77–100.

Calem, Paul, Francisco Covas, and Jason Wu. 2013. "The Impact of the 2007 Liquidity Shock on Bank Jumbo Mortgage Lending." *Journal of Money, Credit and Banking* 45(s1)*:* 59–91.

Campbell, John Y., Stefano Giglio, and Parag Pathak. 2011. "Forced Sales and House Prices." *American Economic Review* 101(5): 2109–31.

DeFusco, Anthony A., and Andrew Paciorek. 2014. "The Interest Rate Elasticity of Mortgage Demand: Evidence From Bunching at the Conforming Loan Limit." Finance and Economics Discussion Series, 2014-11, Board of Governors of the Federal Reserve System.

Di Maggio, Marco, Amir Kermani, and Rodney Ramcharan. 2014. "Monetary Policy Pass-Through: Household Consumption and Voluntary Deleveraging." Columbia Business School Research Paper No. 14-24, November. Available at SSRN.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91(3): 401–19.

Diamond, Douglas W., and Raghuram G. Rajan. 2011. "Fear of Fire Sales, Illiquidity Seeking, and Credit Freezes." *Quarterly Journal of Economics* 126(2): 557–91.

Dwyer, Gerald P., and Paula Tkac. 2009. "The

Financial Crisis of 2008 in Fixed-Income Markets." *Journal of International Money and Finance* 28(8): 1293–1316.

Eberly, Janice, and Arvind Krishnamurthy. 2014. "Efficient Credit Policies in a Housing Debt Crisis." *Brookings Papers on Economic Activity,* Fall.

Eisenbeis, Robert A., W. Scott Frame, and Larry D. Wall. 2007. "An Analysis of the Systemic Risks Posed by Fannie Mae and Freddie Mac and an Explanation of the Policy Options for Reducing Those Risks." *Journal of Financial Services Research* 31(2–3): 75–99.

Fannie Mae. 2012. "Fannie Mae's Analysis Regarding Principal Forgiveness and Treasury's HAMP Principal Reduction Alternative (HAMP PRA) Program." Response to FHFA Request. Washington, DC.

First American CoreLogic. 2010. "Media Alert: Underwater Mortgages On the Rise According to First American CoreLogic Q4 2009 Negative Equity Data." Press Release, February 23.

Foote, Christopher L., Kristopher Gerardi, and Paul S. Willen. 2008. "Negative Equity and Foreclosure: Theory and Evidence." *Journal of Urban Economics* 64(2): 234–45.

Frame, W. Scott, and Lawrence J. White. 2004. "Regulating Housing GSEs: Thoughts on Institutional Structure and Authorities." Federal Reserve Bank of Atlanta *Economic Review* 89(2): 87–102.

Frame, W. Scott, and Lawrence J. White. 2005. "Fussing and Fuming over Fannie and Freddie: How Much Smoke, How Much Fire?" *Journal of Economic Perspectives* 19(2): 159–84.

Fuster, Andreas, Laurie Goodman, David Lucca, Laurel Madar, Linsey Molloy, and Paul Willen. 2013. "The Rising Gap Between Primary and Secondary Mortgage Rates." *Economic Policy Review* 19(2): 17–39.

Fuster, Andreas, and James Vickery. 2015. "Securitization and the Fixed-Rate Mortgage." *Review of Financial Studies* 28(1): 176–211.

Fuster, Andreas, and Paul S. Willen. 2012. "Payment Size, Negative Equity, and Mortgage Default." Staff Report 582, Federal Reserve Bank of New York, November. http://www.newyorkfed.org/research/staff_reports/sr582.pdf.

Gagnon, Joseph, Matthew Raskin, Julie Remache, and Brian Sack. 2011. "The Financial Market Effects of the Federal Reserve's Large-Scale Asset Purchases." *International Journal of Central Banking* 7(1): 3–43.

Geithner, Timothy F. 2014. *Stress Test: Reflections on Financial Crises.* Crown Publishing Group.

Goodman, Laurie S., and Jun Zhu. 2013. "Reps and Warrants: Lessons From the GSEs Experience." Housing Finance Policy Center Working Paper, Urban Institute, October 24. http://www.urban.org/UploadedPDF/412934-Reps-and-Warrants-Lessons-from-the-GSEs-Experience.pdf.

Greenspan, Alan. 2004. "Government Sponsored Enterprises." Testimony before the Committee on Banking, Housing and Urban Affairs, February 24, 2004.

Greenspan, Alan. 2005. "Government Sponsored Enterprises," Remarks to the Conference on Housing, Mortgage Finance, and the Macroeconomy, Atlanta, GA, May 19, 2005 (via satellite). Federal Reserve Board, Federal Reserve Bank of Atlanta.

Hancock, Diana, and Wayne Passmore. 2011. "Did the Federal Reserve's MBS Purchase Program Lower Mortgage Rates?" *Journal of Monetary Economics* 58(5): 498–514.

Haughwout, Andrew, Ebiere Okah, and Joseph Tracy. 2010. "Second Chances: Subprime Mortgage Modification and Re-Default." Staff Report No. 417, Federal Reserve Bank of New York, August.

Hubbard, R. Glenn. 2003. "Evaluating Liquidity Risk Management at Fannie Mae." *Fannie Mae Papers* 2(5): 1–12.

Kaufman, Alex. 2014. "The Influence of Fannie and Freddie on Mortgage Loan Terms." *Real Estate Economics* 42(2): 472–96.

Keys, Benjamin J., Tomasz Piskorski, Amit Seru, and Vincent W. Yao. 2014. "Mortgage Rates, Household Balance Sheets, and the Real Economy." NBER Working Paper 20561, October.

Kiyotaki, Nobuhiro, and John Moore. 1997. "Credit Cycles." *Journal of Political Economy* 105(2): 211–48.

Krishnamurthy, Arvind. 2010. "How Debt Markets Have Malfunctioned in the Crisis." *Journal of Economic Perspectives* 24(1): 3–28.

Krishnamurthy, Arvind, and Annette Vissing-Jorgensen. 2011. "The Effects of Quantitative Easing on Interest Rates: Channels and Implications for Policy." *Brookings Papers on Economic Activity,* Fall, 215–65.

Kung, Edward. 2014. "The Effect of Credit Availability on House Prices: Evidence from the Economic Stimulus Act of 2008." November 30. https://sites.google.com/site/edwardkung/k_cll_2014nov.pdf.

Leitner, Yaron. 2011. "Why Do Markets Freeze?" Federal Reserve Bank of Philadelphia *Business Review,* 2nd Quarter, 12–19.

Lockhart, James B. 2009. "FHFA's First Anniversary and Challenges Ahead." Speech, July 30. http://www.fhfa.gov/Media/PublicAffairs/Pages/FHFAs-First-Anniversary-and-Challenges-Ahead-at-the-National-Press-Club.aspx.

Mayer, Christopher, Karen Pence, and Shane M. Sherlund. 2009. "The Rise in Mortgage Defaults." *Journal of Economic Perspectives* 23(1): 27–50.

McKenzie, Joseph A. 2002. "A Reconsideration of the Jumbo/Non-Jumbo Mortgage Rate

Differential." *Journal of Real Estate Finance and Economics* 25(2–3): 197–213.

Mian, Atif, and Amir Sufi. 2014. *House of Debt: How They (and You) Caused the Great Recession, and How We Can Prevent It from Happening Again.* University of Chicago Press.

Nothaft, Frank, James E. Pearce, and Stevan Stevanovic. 2002. "Debt Spreads between GSEs and Other Corporations." *Journal of Real Estate Finance and Economics* 25(2–3): 151–72.

Passmore, Wayne. 2005. "The GSE Implicit Subsidy and the Value of Government Ambiguity." *Real Estate Economics* 33(3): 465–86.

Paulson, Jr., Henry M. 2010. *On the Brink: Inside the Race to Stop the Collapse of the Global Financial System.* New York, NY: Business Plus.

Philippon, Thomas, and Vasiliki Skreta. 2012. "Optimal Interventions in Markets with Adverse Selection." *American Economic Review* 102(1): 1–28.

RealtyTrac. 2014. "All-Cash Share of U.S. Home Sales Pulls Back from 3-Year High, Institutional Investor Share Drops to 3-Year Low." August 18. http://www.realtytrac.com/content/foreclosure-market-report/q2-2014-us-institutional-investor-and-cash-sales-report-8126.

Rice, Tara, and Jonathan Rose. 2012. "When Good Investments Go Bad: The Contraction in Community Bank Lending After the 2008 GSE Takeover." International Finance Discussion Papers No. 2012-1045, Board of Governors of the Federal Reserve System, March. http://www.federalreserve.gov/pubs/ifdp/2012/1045/default.htm.

Scharfstein, David, and Adi Sunderam. 2014. "Market Power in Mortgage Lending and the Transmission of Monetary Policy." September. http://people.hbs.edu/asunderam/Mortgage%20Market%20Power%2020140907.pdf.

Shleifer, Andrei, and Robert W. Vishny. 1992. "Liquidation Values and Debt Capacity: A Market Equilibrium Approach." *Journal of Finance* 47(4): 1343–66.

Stiglitz, Joseph, Jonathan M. Orszag, and Peter R. Orszag. 2002. "Implications of the New Fannie Mae and Freddie Mac Risk-Based Capital Standard." *Fannie Mae Papers* 1(2): 1–10.

Swagel, Phillip. 2009. "The Financial Crisis: An Inside View." *Brookings Papers on Economic Activity*, Spring, 1–63.

Tracy, Joseph, and Wright, Joshua. 2012. "Payment Changes and Default Risk." Staff Report 562, Federal Reserve Bank of New York, June. http://www.newyorkfed.org/research/staff_reports/sr562.pdf.

US Congressional Budget Office. 2010. "CBO's Budgetary Treatment of Fannie Mae and Freddie Mac." Background paper. January.

US Department of Treasury. 2012. "Treasury Department Announces Further Steps to Expedite Wind Down of Fannie Mae and Freddie Mac." August 17, 2012. Washington, DC. http://www.treasury.gov/press-center/press-releases/Pages/tg1684.aspx.

US Department of Treasury and the Department of Housing and Urban Development. 2011. "Reforming America's Housing Finance Market: A Report to Congress." Washington, DC.

US Federal Housing Finance Agency. 2008. "Statement of FHFA Regarding Contracts of Enterprises in Conservatorship." Press statement, September 7. http://www.fhfa.gov/Media/PublicAffairs/Pages/Statement-of-FHFA-Regarding-Contracts-of-Enterprises-in-Conservatorship.aspx.

US Federal Housing Finance Agency. 2011. "Conservator's Report on the Enterprises' Financial Performance, Fourth Quarter 2011."

US Federal Housing Finance Agency. 2013. "Conservator's Report on the Enterprises' Financial Performance, First Quarter 2013." http://www.fhfa.gov/AboutUs/Reports/Pages/FHFA-Conservator%27s-Report-on-the-Enterprises%27-Financial-Performance---First-Quarter-2013-.aspx.

US Federal Housing Finance Agency. 2014. *Report to Congress 2013*, June 13.

US Federal Housing Finance Agency, Office of Inspector General. 2012. "Fannie Mae and Freddie Mac: Where the Taxpayers' Money Went." White Paper: WPR-2012-02, May 24.

US Financial Crisis Inquiry Commission. 2011."The Financial Crisis Inquiry Report." January. http://fcic-static.law.stanford.edu/cdn_media/fcic-reports/fcic_final_report_full.pdf.

US General Accounting Office [Now, the Government Accountability Office]. 1990. *Government-Sponsored Enterprises: The Government's Exposure to Risks*, August 15. Washington, DC. http://www.gao.gov/products/GGD-90-97.

Vickery, James, and Wright, Joshua. 2013. "TBA Trading and Liquidity in the Agency MBS Market." *Economic Policy Review* 19(1): 1–18.

Wall, Larry D. 2014. "Have the Government-Sponsored Enterprises Fully Repaid the Treasury?" Center for Financial Innovation and Stability, Federal Reserve Bank of Atlanta, March. http://www.frbatlanta.org/cenfis/pubscf/nftv_1403.cfm.

Wall, Larry D., Robert A. Eisenbeis, and W. Scott Frame. 2005. "Resolving Large Financial Intermediaries: Banks versus Housing Enterprises." *Journal of Financial Stability* 1(3): 386–425.

Zhu, Jun, Jared Janowiak, Lu Ji, Kadiri Karamon, and Douglas A. McManus. Forthcoming. "The Effect of Mortgage Payment Reduction on Default: Evidence from the Home Affordable Refinance Program." *Real Estate Economics.*

Journal of Economic Perspectives—Volume 29, Number 2—Spring 2015—Pages 53–80

An Assessment of TARP Assistance to Financial Institutions

Charles W. Calomiris and Urooj Khan

How should economists and policymakers evaluate the assistance provided to financial institutions during the recent financial crisis, and in particular the assistance provided through the 2008 Troubled Asset Relief Program, commonly known as TARP? We examine that question in five parts: 1) What did policymakers do? 2) What are the proper objectives of interventions like TARP assistance to financial institutions? 3) Did TARP succeed in those economic objectives? 4) Were TARP funds allocated purely on an economic basis, or did political favoritism play a role? 5) Would alternative policies, either alongside or instead of TARP, and alternative design features of TARP, have worked better?

In assessing the TARP, we distinguish between the assistance provided to very large banks and that provided to other banks. The largest banks were treated very differently: they were pressured to participate in the initial TARP program, and some were also pressured to participate (through stress testing) in various second-stage programs. Furthermore, the second-stage investments made into these large institutions (which were justified by a belief that these institutions were special because they were "too big to fail") sometimes took very different and riskier forms from the preferred stock and warrant investments made in other banks under the first phase of TARP.

TARP was not a single approach to assisting weak banks but rather a variety of changing solutions to a set of evolving problems. Understanding and evaluating it as

■ *Charles Calomiris is the Henry Kaufman Professor of Financial Institutions in the division of finance and economics at Columbia Business School. Urooj Khan is an assistant professor in the accounting division at Columbia Business School. Their email addresses are cc374@columbia.edu and uk2117@columbia.edu.*

http://dx.doi.org/10.1257/jep.29.2.53 doi=10.1257/jep.29.2.53

such produces a healthy respect for the political constraints that bailout programs face and also points to shortcomings in the ways economists account for the costs of such programs. The political constraints that TARP confronted limited its structure and effectiveness and encouraged it to employ implicit options as a means of assistance, which made the costs of TARP assistance higher than conventional cost calculations have recognized.

Six years after the passage of TARP, it remains hard to measure the total social costs and benefits of the assistance to banks provided under TARP programs. TARP's passage was associated with significant improvements in financial markets and the health of financial intermediaries, as well as an increase in the supply of lending by recipients. However, a full evaluation must also take into account other factors: the risks borne by taxpayers in the course of the bailouts; moral-hazard costs that could result in more risk-taking in the future; and social costs related to the perceived unfairness of the bailouts and the evidence of corruption in the administration of TARP. These effects are difficult to measure. In addition, the TARP experience offers some lessons about how best to assist financial institutions when such assistance is deemed necessary. Going forward, it may be advisable to design a bank assistance program in advance so that its design features can reflect more thoughtful and less politicized judgments about optimal structure and about the social costs and benefits of mitigating systemic risk in the banking system.

The Crisis of 2007–2009 and the Creation of TARP Assistance for Financial Institutions

Policymakers initially responded to the financial crisis in late 2007 and into 2008 with various emergency initiatives: for example, new Federal Reserve lending facilities for banks and other financial institutions; Fed-assisted bailouts of the investment bank Bear Stearns in March 2008; the conservatorship and Treasury "bazooka" bailout of Fannie Mae and Freddie Mac in the summer of 2008; and the bailout of the insurance company AIG in September 2008.[1] The decision in September 2008 not to bail out another investment bank, Lehman Bros., coincided with the continuing deepening of the crisis, which was visible in the price declines suffered by risky assets and bank stocks. That deepening reflected a process of ongoing learning about the extent to which many financial institutions held positions related to deeply troubled assets—"subprime" and "Alt-A" mortgages and the securities backed by them.

By late September 2008, market prices for the shares of the largest banks, including Citigroup, Goldman Sachs, Morgan Stanley, and JPMorgan Chase, had fallen dramatically. The implied market equity ratios (the ratio of market value of equity to the market value of assets) of these banks had fallen so much that

[1] For an overview of the financial crisis of 2007–2009 and the various government responses to it, see Calomiris, Eisenbeis, and Litan (2011).

in some cases those ratios indicated market perceptions of potential insolvency (Calomiris and Herring 2013). As perceptions of default risk rose, banks found it hard to roll over their uninsured debts. Amounts and maturities shrank in markets involving overnight lending between large banks, like the federal funds and LIBOR (London Interbank Offered Rate) markets, and banks hoarded increasing amounts of cash (Heider, Hoerova, and Holthausen forthcoming; Gorton and Metrick 2012; Covitz, Liang, and Suarez 2013).

Amidst this turmoil, as the net worth of banks plummeted, some of the largest financial institutions succumbed to failure or acquisition, and the surviving ones scrambled to pay off maturing debts and restore confidence. Federal Reserve and Treasury officials became convinced that a systematic approach to financial system solvency risk was needed—not just expanded Fed lending programs and bailouts in response to some individual failures—to maintain confidence in the financial system and to ensure that banks continued to supply loans and other essential financial needs of the economy.

Treasury Secretary Henry Paulson and Fed Chairman Ben Bernanke testified numerous times together before Congress in mid- to late-September 2008 in favor of shoring up the banking system with additional measures to prevent a systemic collapse. Paulson proposed government assistance to banks in the form of support for selling troubled mortgage-related assets at prices that were more reflective of their long-term earnings potential, which he argued were far in excess of their current prices. The discussion in Congressional hearings of options for assistance was narrowly confined to the Secretary's proposal; independent voices with alternative views on whether or how to provide systemic assistance to the banking system were not invited to testify before Congress in the weeks it deliberated over TARP.[2]

Secretary Paulson appeared repeatedly to defend what became known as the Troubled Asset Relief Program (TARP). It took about three weeks for Congress to approve TARP. Some of the initial Congressional resistance to the bailout plan was eroded by the adverse stock market reaction to the failure to win passage of TARP on September 29. On October 3, 2008, the Emergency Economic Stabilization Act (EESA) of 2008, which established up to $700 billion (outstanding at any one time) in TARP assistance, passed both houses of Congress and was signed by President Bush. On October 13, the Treasury announced a new plan to invest in bank capital via the Capital Purchase Program (CPP). On October 14, nine large

[2] Some alternatives were proposed, including Senator Charles Schumer's proposal, presented in a mid-September speech, in which he advocated the use of bank preferred stock purchases by the government alongside mortgage relief for homeowners. Schumer referenced the 1933 preferred stock purchases of the Reconstruction Finance Corporation. In his follow-up op-ed in the *Wall Street Journal* (October 14, 2008), he also advocated the prohibition of common stock dividends to banks receiving government preferred stock assistance, and for providing assistance in a way that would "encourage private investors to make similar investments." These proposals echoed the views of some academic policy advocates, including one of us (Calomiris 2008). Not all members of Congress were receptive to the shift in TARP from asset purchases to the capitalization of banks; the US Government Accountability Office (2009, p. 10) describes the reaction as a "backlash" and used it to support its recommendations of enhanced transparency and communications throughout its early oversight of TARP.

financial institutions (under the coordination and reportedly also the pressure of the Treasury), which together accounted for 55 percent of US banks' assets, announced that they would subscribe for a total of $125 billion of TARP assistance (GAO 2012a, p. 7). The nine institutions were Bank of America, Citigroup, JP Morgan Chase, Wells Fargo, Morgan Stanley, Goldman Sachs, Bank of New York Mellon, State Street, and Merrill Lynch. Other publicly traded financial institutions were eligible to apply until November 14, 2008 (all of which presumably participated on a purely voluntary basis).

Secretary Paulson's initial vision of TARP was a mechanism through which the government would support the sale of the "troubled" assets of banks to the government through a complex process, or by having the government guarantee the value of the assets at prices in excess of crisis-affected market values. By raising the asset values of banks, TARP would restore market confidence in bank solvency, and allow debt and lending markets to be restored to normalcy. But the Treasury soon abandoned that approach in favor of direct government injections of capital into banks in the form of preferred stock purchases. Preferred stock purchases had been authorized under TARP almost as an afterthought; indeed, the authority for purchases of bank preferred stock is a bit hard to discern from reading the statute. Any purchases of securities (such as preferred stock) had to be accompanied by the granting of warrants (which allow future purchases of stock from the firm at a pre-established price) to ensure that taxpayers shared in the upside potential of recipient institutions, and those warrants should also include anti-dilution provisions "of the type employed in capital market transactions."

TARP's Conflicting Goals and Constraints

Although the first stated purpose for TARP (under Section 2 of the Act) was "to immediately provide authority and facilities that the Secretary of the Treasury can use to restore liquidity and stability to the financial system of the United States," its other stated purpose was "to ensure that such authority and such facilities are used in a manner that—(A) protects home values, college funds, retirement accounts, and life savings; (B) preserves homeownership and promotes jobs and economic growth; (C) maximizes overall returns to the taxpayers of the United States; and (D) provides public accountability for the exercise of such authority."

Items (A) and (B) presented special challenges, especially if the Treasury acquired troubled assets through direct asset purchases under Section 101 of the law. Any acquisition of mortgages or mortgage-backed securities by the Treasury would put it in the position of having to determine the extent of relief to homeowners, which would require weighing the direct financial costs to taxpayers against the benefits to homeowners and the economy (and the consequent indirect benefits to taxpayers). Under Section 109, the Secretary was charged with implementing a plan that both "seeks to maximize assistance for homeowners" while "considering net present value to the taxpayer." No wonder the Treasury opted to abandon direct asset purchases. Not only was it impossible to establish fair prices for such assets, but doing so would have put Treasury directly in charge of mortgage restructuring, while facing an impossible mandate to meet an amorphous objective of "maximizing assistance" while minimizing costs to taxpayers.

The constraints contained in items (C) and (D) of Section 2 were also serious, and they applied to all forms of TARP assistance. In reaction to Lehman's failure, Warren Buffett had just purchased a substantial amount of Goldman Sachs preferred stock and had received warrants to purchase equity in addition to the promised coupon payments on the preferred stock. Item (C) seems to have been intended in part to ensure that taxpayers' investments in preferred stock were treated as similarly profit-making investments. Purchases of assets under TARP were supposed to be priced to maximize taxpayers' returns (broadly defined). Government guarantees of assets under Section 102(c) were even more constrained by an explicit requirement to earn an actuarially fair market insurance premium. TARP also included limits on executive compensation, designed to prevent profiteering from government assistance (especially with respect to golden parachutes for executives), and those compensation limits were tightened over time.

The Emergency Economic Stabilization Act of 2008, which established TARP, did not require that purchases of preferred stock assistance be provided on market terms, as it allowed the Secretary of the Treasury, under Section 113(a), when minimizing the "long-term negative impact on the taxpayer" to take into account not only "the direct outlays, [and the] potential long-term returns on assets purchased," but also "the overall economic benefits due to improvements in economic activity and the availability of credit, the impact on the savings and pensions of individuals, and reductions in losses to the Federal Government." In other words, the Secretary was told to take into account the positive externalities taxpayers accrued through expanded credit and economic activity.

TARP took the unusual step of requiring the Office of Management and Budget (OMB) and the Congressional Budget Office (CBO) to perform a *true economic cost accounting* for TARP (under Section 202) that "shall be calculated by adjusting the discount rate . . . for market risks" (Section 123). The conclusions of that accounting had to be included in federal budgetary accounts as supplementary materials (Section 203). In other words, any subsidies provided to banks would be explicitly estimated using economic measures of opportunity cost, and under Section 113(a), it would be the obligation of the Secretary of Treasury to ensure that indirect benefits to taxpayers equaled or exceeded those costs.

In this politicized environment, operating under these conflicting and unclear mandates, the Treasury focused on preferred stock purchases. Doing so allowed it to avoid the zero-subsidy constraint applicable to asset guarantees and the potential problems associated with buying troubled mortgages at defensibly fair prices and managing them under the conflicting mandates of the law. As of the end of 2009, a total of 707 financial institutions received a total of $205 billion under the Capital Purchase Program.

The Treasury set uniform terms for preferred stock purchases under the Capital Purchase Program, requiring a 5 percent initial coupon on preferred stock, rising to 9 percent after five years, and demanding 15 percent of preferred stock infusions be in the form of 10-year warrants to purchase common stock. It limited participation to "qualifying" banks, which in practice meant banks that were not so deeply troubled that they were likely to fail even after receiving preferred stock assistance. Investments under the CPP initially were limited to between 1 and 3 percent of a bank's risk-weighted assets and were capped at $25 billion (US GAO 2012a, p. 4).[3]

Although the banks may have felt the Treasury's preferred stock investment terms were expensive, the terms Warren Buffett negotiated with Goldman Sachs for Berkshire Hathaway, in a deal announced on September 23, 2008, allowed Berkshire an even higher return. Berkshire Hathaway, had received 100 percent of the $5 billion preferred stock issue in warrants with a five-year term, and a 10 percent coupon on the preferred stock. The Goldman Sachs preferred stock offered to Berkshire was callable at any time at a 10 percent premium.[4]

Government preferred stock purchases required participating issuers to freeze their common stock dividends, but issuers were not forced to shrink dividends as a requirement for participating in the Capital Purchase Program (implying that recipient banks were effectively able to subordinate preferred stock through the payment of common stock dividends). Limits on dividends have been shown to be very useful in limiting abuse of government protection (Calomiris and Mason 2004; Hovakimian, Kane, and Laeven 2012), but these limits reportedly were not feasible in light of the desire to encourage all large banks (including those not in need of the assistance) to participate. Secretary Paulson effectively forced the largest US banks to participate in the CPP (Veronesi and Zingales 2010; Kim and Stock 2012), and those that did not need the assistance balked at any limit on their dividends. Paulson may have agreed to permit the continuing payment of common stock dividends in order to achieve the policy goal of uniform participation, arguably a symbolic victory.

Phase Two: The SSFI, AGP, CAP, and TIP Programs

After the 2008 election, TARP assistance changed. Attention turned to evaluating and addressing the circumstances of particular large institutions whose financing structure remained problematic, and the nature of assistance was more varied. Although funding through the Capital Purchase Program continued, new sources of funding were designed to deliver customized assistance, alongside the more general approach. The four parts of the second phase included: the Systemically Significant Failing Institutions (SSFI) Program, the Asset Guarantee Program (AGP), the Targeted Investment Program (TIP), and the Capital Assistance Program (CAP).

[3] In May 2009, this provision was amended so that qualifying financial institutions with total assets less than $500 million would receive investments between 3 and 5 percent of risk-weighted assets.

[4] In fact, the preferred stock was called by Goldman Sachs in March 2011. Rather than exercising its warrants, Berkshire ended up making a settlement in March 2013, exchanging its warrants for roughly 13 million shares of Goldman Sachs common stock (2.8 percent of the company). All told, from September 2008 to March 2013, Berkshire Hathaway made roughly $3.7 billion in income on its $5 billion initial investment in preferred shares Information about the Berkshire Hathaway purchase of Goldman Sachs securities is from Goldman Sachs (2008). Returns on this investment are based on various news stories and on authors' calculations.

The SSFI, AGP, and TIP were created to meet the needs for what the Treasury termed "exceptional assistance" by three institutions: AIG, Citigroup, and Bank of America.

Assistance remained controversial during this second phase of TARP, and growing public resentment over high compensation in assisted banks led to stricter limits on executive compensation for TARP recipients. This not only resulted in greater reluctance of banks to apply for TARP funding, it also resulted in substantial repurchases of preferred stock as a means of exiting from the discipline of the increasingly stringent compensation regulations that were attached to government investments.

By the end of 2009, $70.7 billion of $204.6 billion disbursed under the Capital Purchase Program had been repurchased by participating banks. Five of the large banks that were among the nine original participants repurchased their CPP securities in June 2009 (GAO 2009, pp. 8, 13). The CPP was closed to new investments at the end of 2009, and as of September 20, 2010, two years after TARP had been passed, the Capital Purchase Program had been largely wound down with $152 billion of investments under that program having been repaid (GAO 2011b, p. 13). Participants that did not exit TARP by 2012 were relatively weak, had larger loan losses, and increasingly displayed problems in paying dividends and maintaining profitability (GAO 2013b, p. 5). In November 2013, the Treasury estimated the eventual nominal gains on all CPP investments would be roughly $16 billion (GAO 2014, pp. 1–5). The program had succeeded in improving banks' capital levels, thereby enhancing their ability to borrow and lend.

The first new program under the post-election phase of TARP was the Systemically Significant Failing Institutions plan, announced on November 10, 2008, to purchase AIG preferred stock (the only use ever made of SSFI; SSFI was later renamed the AIG Investment Program). The AIG situation is discussed in the paper by Robert McDonald and Anna Paulson in this symposium. Total Treasury and Fed exposure to AIG reached an astounding $172.4 billion at the end of 2009—nearly equal to the entire amount disbursed under the Capital Purchase Program. Its form changed over time from relatively senior obligations (preferred stock) to junior ones (common stock). The changing structure of that assistance is so complex that it took a 70-page report by the General Accountability Office just to describe the program's evolution. On December 14, 2012, the Treasury announced that it had received the proceeds from its final sale of AIG stock, ending the government's complex program of assistance to AIG, and resulting in a slight income of $2.3 billion over its funds invested in AIG (US GAO 2013a, p. 5).

Citigroup was the only financial institution to participate in the Treasury's Asset Guarantee Program, although Bank of America also considered participating. On January 15, 2009, Citigroup arranged for loss protection on a $301 billion portfolio of assets, which created a potential exposure of $5 billion for the Treasury, and paid for that protection with preferred shares and warrants. Over its lifetime, the total net income the Treasury gained under this guarantee program was $3.9 billion.

Citigroup and Bank of America were the only banks to receive assistance under the Targeted Investment Program, under agreements finalized, respectively, on

December 31, 2008, and on January 15, 2009. Under TIP, the Treasury invested $20 billion in each and received preferred stock and warrants. TIP imposed looser standards for approval than the Capital Purchase Program and was directed toward banks with special systemic importance. Consistent with the targeted nature of this assistance, receiving TIP assistance was also associated with "stringent regulations regarding executive compensation, lobbying expenses, and other corporate governance requirements" (US GAO 2009, p. 73). The Treasury's TIP investment in Citigroup was converted into common stock in September 2009. The ultimate recoveries from the various TIP-related investments exceeded the cost basis of Treasury TIP investments by $4.0 billion (GAO 2013a, p. 5).

Treasury Secretary Timothy Geithner assumed office under the Obama administration in January 2009 and initiated a Financial Stability Plan, which established new stress tests to gauge the fragility of the largest banks and linked TARP assistance to the results of those stress tests. On February 17, 2009, Title VII of the American Recovery and Reinvestment Act (ARRA) amended the Emergency Economic Stabilization Act of 2008 to establish new compensation rules for TARP assistance to financial institutions and to permit those that had received Capital Purchase Program assistance to buy back preferred stock and warrants with the approval of their regulators. The Capital Assistance Program was established February 25, 2009, mandating that banks with assets in excess of $100 billion accept government injections of capital (issuing preferred stock convertible into common stock) if privately raised capital proved inadequate in light of new forward-looking loss assessments usually called the "stress tests." Banks that had previously received CPP assistance were permitted to convert those issues into the new convertible preferred shares.

Under the Capital Assistance Program, it was announced on May 7, 2009, that 10 of the 19 banks subjected to stress tests needed to raise additional capital (of approximately $75 billion in total). They were given six months to do so privately; if they were unable to do so, they had to accept government injections of convertible preferred stock to cover the gap identified by the stress test. Setting up a contingent source of government funding ensured that markets would not be rattled too much by any announced deficiencies, which also made the stress tests more credible as an exercise, as regulators would be more likely to honestly identify deficiencies if doing so was unlikely to roil markets.

No funds were actually disbursed under the Capital Assistance Program, and the program was terminated in November 2009, but the capital deficiencies identified by the May 7, 2009, stress test announcement did produce additional capital raising in private markets and also were associated with major restructuring of the Treasury's investment in Citigroup. In June 2009, Citigroup and Treasury agreed to swap $20 billion in cumulative perpetual preferred stock (issued under the Targeted Investment Program and the Asset Guarantee Program) for a form of preferred stock (so-called trust preferred securities) that counts for regulatory purposes as providing more protection to deposits than other preferred stock, which had the effect of raising Citigroup's tier-1 capital ratio. Citigroup also agreed to swap $25 billion in its Capital Purchase Program preferred stock for an equal amount

Table 1
Cumulative Income by Program, 2008–2013
($billions)

Program	*Maximum exposure*	*Income*[a]
Capital Purchase Program (CPP)	204.6	16.0
Systemically Significant Failing Institutions (SSFI)/AIG[b]	172.4	15.0
Asset Guarantee Program (AGP)	5.0	3.9
Targeted Investment Program (TIP)	40.0	4.0
Total	**422.0**	**38.9**
Total for only Citigroup and AIG	222.4	28.4
Total subtracting Citigroup and AIG	199.6	10.5

Sources: US Government Accountability Office (various).
[a] Cumulative income on CPP includes estimates on income and losses expected for outstanding investments.
[b] Includes some non-TARP programs.

of various interim securities, which were converted into common stock shares on September 3, 2009, making the US government a major junior stakeholder in Citigroup. The Treasury Department sold its common stock in Citigroup in 2010, with the last of those sales completed in December 2010. It auctioned its Citigroup warrants in January 2011, and liquidated the last of its Citigroup-related securities (subordinated notes it had received from the Federal Deposit Insurance Corporation in 2012 as part of the compensation for Citigroup's Asset Guarantee Program coverage) on February 4, 2013. All told, the Treasury received $58.4 billion from its $50 billion investments in Citigroup.[5]

How "Junior" Was Born: Bagehot's Rule Meets "Too-Big-To-Fail"

During the post-election phase of TARP, common stock became an important part of the Treasury's portfolio of investments in financial institutions. Interestingly, the returns earned on the common stock investments in AIG and Citigroup were similar to the returns on the Capital Purchase Program investments made in other financial institutions. As Table 1 shows, total cumulative income on investments in AIG and Citigroup were 12.8 percent of maximum exposures ($28.4 billion relative to $222.4 billion), while the income on the remaining investments (which did not include common stock) were only 5.3 percent of maximum exposures ($10.5 billion relative to $199.6 billion). On an annualized basis, the returns for these two subsets of investments were similar, reflecting the fact that the durations of the Citigroup and AIG common stock investments were longer than the roughly one-year average

[5] The Treasury improperly refers to its return relative to a $45 billion investment in Citigroup, which omits its $5 billion of loss exposure on the AGP program. For the details of the timing of the various Treasury sales of Citigroup's shares, warrants, and debt, see Braithwaite and Guerrea (2010), Griffen (2011), and US Treasury (N.d.).

duration of the portfolio of CPP investments in other banks. The duration of the Treasury's investments in Citigroup were more than two years, and the average duration of the government's investments in AIG was even longer. However, neither of these returns compares favorably with Berkshire Hathaway's 74 percent cumulative return over 4.5 years on its preferred investment in Goldman Sachs.

Of course, the success of TARP should not be measured solely or even primarily on the basis of realized returns. Realized returns on common stock investments generally should be higher than realized returns on preferred stock investments, but in the case of TARP, that was not true because investments in common stock were made *selectively*. Preferred stock and debt investments were converted into common stock in Citigroup and AIG precisely because of the continuing weak financial condition of these firms in 2009 and 2010. Thus, it is no surprise that realized returns on their common stock were meager. In other words, any TARP investment in a too-big-to-fail bank *had always been* an implicit contingent common stock investment, which would convert to common stock as needed to preserve the "too-big-to-fail" institution. It was unlikely that the government would use its preferred status in the states of the world where it would be financially useful to do so (in bankruptcy or receivership) because the government would convert to common stock in order to prevent bankruptcy or receivership.

This contingent equity aspect of TARP investments in too-big-to-fail institutions highlights one of the respects in which TARP differed from conventional debt or preferred stock programs of bank assistance like, for example, collateralized lending by a central bank under "Bagehot's Rule," or the Reconstruction Finance Corporation's (RFC) preferred stock program initiated in March 1933.[6] Collateralized lending to banks relies upon the use of relatively high-quality assets to make government loans less risky to the central bank or taxpayers. This form of assistance can be effective in resolving pure liquidity problems (where banks lack cash but their problems do not reflect a significant increase in their risk of insolvency). Collateralized lending does not work, however, when bank illiquidity is a symptom of substantially increased default risk of the bank. In such circumstances, the use of collateralized lending can actually exacerbate the liquidity problems of a bank by effectively subordinating the bank's depositors to the central bank or government lender (as depositors' claims become effectively junior to the new lender and are backed by relatively risky assets). Under such circumstances, a collateralized loan that raises the riskiness of deposits might even cause a depositor run rather than prevent one.

With that specific problem in mind, the Roosevelt administration implemented a preferred stock program for assistance to financial institutions as part of the Emergency Banking Relief Act of March 9, 1933. Investments of preferred stock were not

[6] For studies of policies of the Reconstruction Finance Corporation and their effects on bank survival and lending see Mason (2001), Calomiris and Mason (2004), Calomiris, Mason, Weidenmier, and Bobroff (2013), and additional references in these studies. On theory of preferred stock as an effective tool, see Philippon and Schnabl (2013).

collateralized, were junior to all bank debt, including deposits, and failure to pay a preferred stock coupon did not force a bank into conservatorship. Thus, preferred stock added protection to deposits. At the same time, preferred stock was senior to common stock, which served as a buffer against losses on assets.

Preferred stock investments in banks, however, are not appropriate for assisting all banks. As fixed income investments that are senior to common stock, they contribute to highly leveraged banks' risk-management incentive problems, which are also known as the "debt overhang" problem (Jensen and Meckling 1976; Myers 1977; Hoshi and Kashyap 2010). The existing shareholders/managers of a bank that is close to insolvent or actually insolvent see little gain to themselves from limiting the risk of bank investments or finding good loan customers that would raise the bank's revenues as reductions in risk or expansions of cash flow would mainly accrue to other (senior) bank claimants. Providing more preferred stock to such a bank will add to its debt overhang problem and further discourage efforts to raise common stock, identify good loan customers, and manage risk properly and therefore may be socially wasteful.[7]

What can the government do when debt overhang makes preferred stock an undesirable means of assistance? One option is to force the bank to become a target in an assisted merger. This approach is often taken by the Federal Deposit Insurance Corporation for undercapitalized or insolvent banks, but it may not be feasible for a large bank given the difficulty in finding a large acquirer quickly (a problem further complicated by concerns about the increased concentration of banking in an already highly concentrated banking system). It is important to emphasize the speed with which resolution of a financial institution should occur. Global banks are counterparties in numerous short-term transactions; in order to avoid disruption to their operations and the operations of their counterparties, a bank must be resolved immediately upon any regulatory intervention that places it into conservatorship. Another option would be to place the bank into receivership and liquidate its assets without trying to find an acquirer. But institutions like Citigroup or AIG were regarded as "too big to fail," owing to their global scope, the complexity of their subsidiary structures, and their widespread linkages throughout the global financial system.

Still another option in the presence of debt overhang would be to purchase the institution's assets at above-market values, or to provide a subsidy to the institution in a way that guarantees those assets' values. Either of those actions would raise the market value of the equity of the institution, thereby alleviating its debt overhang problem. In a similar vein, the government could attach guarantees (effectively offering a put option) to public offerings of common stock issues by the institution,

[7] The debt-overhang problem can be solved in some cases by requiring issues of subsidized preferred stock to be matched by new common stock issues (Calomiris 1998, 2008). However, when banks are in a very severe debt overhang situation, the ability to offer subsidies on preferred stock to encourage such matching is limited by the zero-coupon bound (the maximum subsidy that can be given for issuing preferred stock), and severely indebted banks may not be willing or able to satisfy such matching requirements.

which would raise the price of those offerings to an extent that would make offerings of new equity appealing to existing shareholders. In a later section, we assess these sorts of interventions. When neither speedy acquisition nor liquidation seem appropriate, and when subsidized put options on assets or new stock offerings are unappealing for some reason, government common equity investments become the path of least resistance for providing assistance to an insolvent, or nearly insolvent, "too-big-to-fail" institution like Citigroup or AIG.

The Objectives of Government Intervention to Assist Financial Institutions

Given the financial costs and design challenges of assisting banks, what prospective benefits may justify such costs? During the Depression, Irving Fisher and John Maynard Keynes articulated various channels through which weak banks can amplify macroeconomic downturns through reduced lending and asset price declines. This thinking became more integrated into macroeconomic thinking (not coincidentally) during the 1980s, particularly as the result of Bernanke's (1983) work on the Great Depression and his and others' empirical work on the macroeconomic consequences of US banks' losses of bank capital in the 1980s (for example, Bernanke and Lown 1991).[8]

Banks are highly leveraged entities that act as repositories of private information about borrowers and securities issuers. Theories of financial intermediation show why their role as information repositories tends to be associated with high leverage (Diamond 1984; Calomiris and Kahn 1991; Krasa and Villamil 1992; Diamond and Rajan 2009). High leverage, however, also means that banks play a central role in propagating economic downturns (Bernanke and Gertler 1989). When shocks to banks' borrowers produce loan losses, some banks fail and survivors' capacity to bear risk declines, forcing cuts in lending.

As Adrian and Shin (2009) show, the real effects of intermediaries' behavior are not confined to declines in lending. Because intermediaries play central roles in asset markets, their shrinkage can have dramatic effects on the prices of risky assets. For example, when hedge funds specializing in emerging market securities

[8] For an early review of the literature on financial factors during the Depression, see Calomiris (1993). Bernanke's (1983) time series study of the links between bank distress and economic activity has been criticized, but subsequent work, using panel data at the level of states or counties, confirms the importance of banking distress as a propagator of shocks during the Depression and also confirms the positive role that assistance to banks via the Reconstruction Finance Corporation played in mitigating the consequences of bank distress (Calomiris and Mason 2003; Calomiris, Mason, Weidenmier, and Bobroff 2013). In addition to the effects of bank condition on lending and securities pricing, Anari, Kolari, and Mason (2005) point to another channel through which bank distress magnified the economic downturn during the 1930s: the protracted process of liquidating the assets of banks that were placed into receivership. Liquidating assets depresses asset values in local markets. Those asset-pricing consequences created an incentive for postponing liquidation, which resulted in protracted delays in depositors' ability to receive repayment of their deposits in failed banks.

lost money during the Russian crisis of 1998, Brazilian international bonds held by these funds were sold off massively. Because other investors not specializing in emerging markets had limited knowledge and consequently limited capacity for bearing emerging market risks, Brazilian sovereign debt prices fell dramatically. These connections between "funding liquidity" of intermediaries and "market liquidity" of securities have been formalized in Brunnermeier and Pedersen (2009).

Many of the debt instruments that banks rely upon for funding require them to maintain near-zero default risk. Because financial intermediaries depend upon risk-intolerant debt instruments (such as interbank deposits, repo, and commercial paper), they are especially vulnerable to adverse shocks to their asset values, which makes shocks to the value of banks' assets (as in the case of subprime mortgages) especially likely to produce sudden declines in credit and in risky asset prices. These channels of transmission were visible in the recent crisis (Gorton 2009; Schwarz 2015; Calomiris 2009a; Heider, Hoerova, and Holthausen forthcoming; Ivashina and Scharfstein 2010; Gorton and Metrick 2012; Covitz, Liang, and Suarez 2013).

If the condition of financial intermediaries is an important propagator of shocks, then it may be useful to shore up the condition of intermediaries as part of a program of combating a recession caused by a major shock to the banking system. There is empirical evidence identifying favorable consequences for lending, asset pricing, and economic activity from assistance to financial intermediaries, policies that seek to improve the financial condition of intermediaries indirectly (for example, through debt re-denominations), or interventions to improve the liquidity of markets in the wake of bank failures (for example, government-sponsored asset management companies).[9] Of course, this argument was used by Paulson and Bernanke in support of Congressional approval of TARP.

The debates over TARP, however, did not *only* reflect economic concerns and arguments, but also other considerations, which affected the process of approving TARP. Deep resentment toward banks—precisely because of their central role in precipitating the crisis—constrained public willingness to assist them. Deep suspicion of government policies to assist banks, which reflected legitimate concerns that government policies may serve special interests rather than the public interest,[10] complicated any attempt by the government to assist banks. Nor was it obvious that government assistance to banks would actually be implemented wisely. For example, it is hard to make sense of the government's decisions to bail out

[9] For a general review, see Calomiris, Klingebiel, and Laeven (2005), who discuss the relative advantages of different policy approaches in different economic environments. See also the aforementioned studies of the operation of the Reconstruction Finance Corporation as a particular example of the effects of preferred stock assistance to banks, and Kroszner (1999) and Calomiris (2007) on the positive macroeconomic consequences of redenomination. Bayazitova and Shivdasani (2012) show that capital injections into banks can be useful as a signal of favorable private information, which can reduce asymmetry of information in public markets.

[10] History confirms that government regulations and government assistance should be understood as political outcomes reflecting the creation of coalitions sufficiently powerful to enact programs, not as the politically neutral application of economic ideas (Calomiris and Haber 2014, chap. 6–8).

Bear Stearns, AIG, and Citigroup, but to refuse to bail out Lehman. Furthermore, it is far from obvious that "too-big-to-fail" bailouts always make sense, especially when one considers the hard-to-measure moral-hazard costs in the future that come from such bailouts today.

The Economic Consequences of TARP

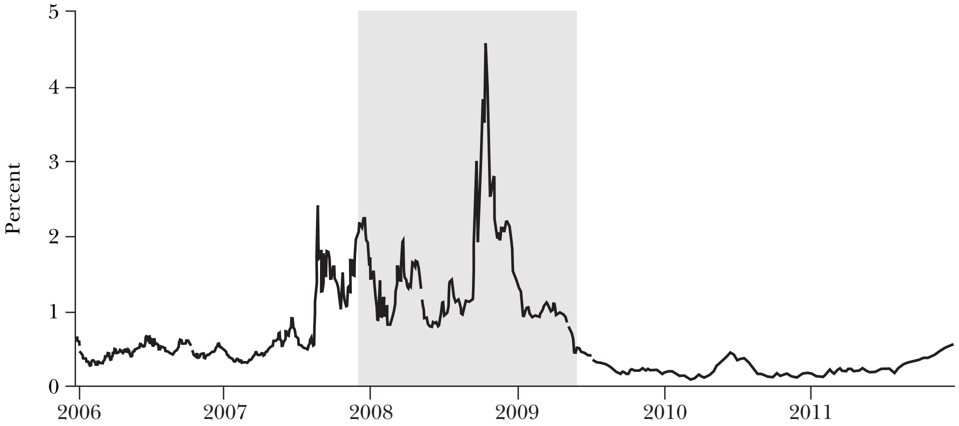
To fulfill TARP's statutory requirements, the Office of Management and Budget and the Congressional Budget Office estimated the costs of TARP's asset purchases and guarantees using procedures similar to those specified in the Federal Credit Reform Act of 1990 with an adjustment for "market risk," as required by the authorizing legislation. The agencies interpreted market risk to be the premium that a private investor would require as compensation for the risk of the cash flows of the underlying transaction. Nominally, there were profits. As of March 12, 2014, the CBO estimated the net cost of TARP to the federal government, measured on the basis of nominal outlays and receipts, to be $27 billion.[11] For the most part, the transactions with the banks, the focus of this paper, yielded a net cash flow gain. The net cash flow costs were largely from the assistance provided to AIG, the automotive industry, and the programs aimed at avoiding home mortgage foreclosures. The net cash flow gain estimated for the Cash Purchase Program was $16 billion with only $2 billion of preferred stock remaining outstanding. The CBO estimated a net cost of $15 billion to the Treasury for the assistance provided to AIG under the Systemically Significant Failing Institutions program. All of the supplementary support provided to Citigroup and Bank of America through the Targeted Investment Program had been paid back and resulted in a net gain of roughly $4 billion dollars to the federal government. Finally, the loss-sharing agreement with Citigroup through the Asset Guarantee Program yielded a net gain of $3.9 billion.

But in evaluating the costs and benefits of TARP, as the authorizing legislation recognized, it is important both to adjust cash flows for the risk borne by taxpayers and to look beyond the net risk-adjusted cash flows received by taxpayers to examine the impact of TARP on the broader economy. After all, the first stated purpose of the program was "to restore liquidity and stability to the financial system of the United States." But measuring risk adjustment on TARP funds (and the implied subsidy received by TARP recipients) and gauging the benefits to the economy from TARP are challenging, to say the least.

The most relevant measure of the subsidy received by TARP recipients is the estimate made at the time the funds were disbursed. The Congressional Budget Office used the market yields on actively traded preferred stock to gauge the size of the subsidy received by preferred stock issuers, and used the Black–Scholes option

[11] The White House Office of Management and Budget estimated the cost of TARP to be $39 billion. The additional estimate of $12 billion from the Congressional Budget Office largely related to CBO's higher projection of costs for the mortgage programs under TARP.

Figure 1
TED Spread



Source: Federal Reserve Bank of St. Louis.
Notes: The TED spread is defined as the difference between the three-month LIBOR and the three-month Treasury bill yield. The shaded area marks the 2007–2009 financial crisis..

pricing model to value warrants. When no preferred stock was available for the issuer, it used a market index. On the first $247 billion of TARP disbursements to banks, the implied subsidy received by program participants, estimated as of the end of 2008, was $64 billion (Congressional Budget Office 2009, p. 1). The Office of Management and Budget's methods for calculating the implied subsidy arrive at comparable numbers. Veronesi and Zingales (2010) calculate a subsidy of between $21 billion and $44 billion on the first $130 billion of TARP disbursements, which implies a comparable proportional value of the subsidy.

One would arrive at a higher subsidy cost estimate if one appropriately recognizes that TARP investments in the largest banks never were just preferred stock. As the experience of Citigroup and AIG show, taxpayers were effectively forced to convert preferred stock to junior equity positions in those institutions because their prospects were slow to improve. In that sense, taxpayers were effectively receiving a fixed income instrument but bearing the risk of losing their senior status on an as-needed basis.

Did the passage of TARP have positive effects on the financial system? Leading up to its passage, market credit spreads had increased to unprecedented levels as investors became increasingly risk-averse due to worries about the health of the banking system and the economy in general. Figure 1 shows the TED spread: that is, the difference between the bank-to-bank overnight lending rate embodied in the London Interbank Offered Rate (LIBOR) and the Treasury bill rate, which captures the extent to which the banking system experienced a crisis of confidence and a reduction in liquidity. The spread increased to 450 basis points, at its highest, in the aftermath of the bankruptcy of Lehman Brothers. Following the announcement of the Capital Purchase Program on October 14, 2008, the first program of TARP announced in the pre-election phase, there were broad improvements in the credit markets. Between Friday, October 10 and Tuesday, October 14, the Standard and Poor's 500 rose by 11 percent and the common stock prices of the nine large financial institutions that were the very first participants of TARP increased by 34 percent (Veronesi and Zingales 2010). From October 13, 2008 (before the announcement of the CPP) to September 30, 2009, the LIBOR rate fell by 446 basis points and TED spread fell by 434 basis points. Costs of credit and perceptions of risk declined significantly in corporate debt markets as well. By the end of September 2009, the Baa bond rate and spread had fallen by 263 and 205 basis points, respectively (US GAO 2009, p. 37).

A specific goal of the Capital Purchase Program was to improve the banks' balance sheets by infusing banks with capital and thereby enhance the ability of banks to borrow and lend. The US Government Accountability Office (2009) reports that capital ratios at institutions that received CPP investments rose more than the ratios at nonparticipating institutions. Between December 31, 2008, and March 31, 2009, the Tier 1 risk-based capital ratio increased by, on average, 300 basis points in bank holding companies receiving CPP assistance relative to an increase of only 40 basis points in nonparticipating bank holding companies. The evidence also suggests that participating banks were more willing and able to increase lending than nonparticipating banks (US GAO 2009; Taliaferro 2009; Ng, Vasvari, and Wittenberg-Moerman forthcoming; Berger and Roman forthcoming; Li 2013). The 21 largest CPP recipients reported extending almost $2.3 trillion in new loans as of July 31, 2009, since receiving CPP investments of $160 billion.

How can one weigh and compare the costs and benefits associated with TARP to arrive at a net benefit estimate? Using an event study analysis of bank enterprise values, Veronesi and Zingales (2010) analyze the effect of the initial announcement of TARP assistance to the financial sector. They estimate that the October 13, 2008, announcement resulted in a net social benefit to financial intermediaries, after subtracting the cost to taxpayers, of between $86 billion to $109 billion, perhaps capturing the benefit of avoiding costly liquidation of financial intermediaries, among other things. This is a lower bound estimate of the social gains from TARP. The authors include in their measure of costs the $125 billion preferred equity infusion in the nine largest US commercial banks via the Capital Purchase Program and a three-year government guarantee on new unsecured bank debt issues provided by the Federal Deposit Insurance Corporation. They find that banks that were more at risk of experiencing a sudden outflow of funding benefited the most from the government's intervention. More specifically, enterprise bank value increased the most for the three former investment banks (Goldman Sachs, Morgan Stanley, and Merrill Lynch) and Citigroup following the October 13 announcements, while the relatively healthy JP Morgan—which stood to gain from the continuing weakening of its troubled rivals—experienced the largest decrease.

The most important limitation of the Veronesi and Zingales (2010) calculation of the net gains from TARP is the authors' assumption that the only anticipated costs to taxpayers under TARP as of October 14, 2008, were the outlays announced under the Capital Purchase Program (and the Federal Deposit Insurance Corporation debt-guarantee). In the event, as initial assistance proved inadequate for Citigroup, AIG, and Bank of America, several more assistance programs were announced by the federal government. To the extent that the potential weakness of these banks was known, and to the extent that the potential additional expenditures in response to that weakness were also forecastable, Veronesi and Zingales (2010) underestimate the expected costs of TARP as of October 13, 2008. The first round of assistance provided to the big banks effectively committed the government to a "whatever it takes" approach to keep AIG, Citigroup, and Bank of America alive, and therefore, the continuing cost to taxpayers actually experienced in 2008–2012 was predictable, at least to some degree. In other words, if TARP assistance would be forthcoming (and more junior in form over time) in response to worsening bank condition, the recipients effectively possessed a put option from the government to issue equity in addition to the explicitly recognized preferred stock investments made by the government. Veronesi and Zingales (2010) do not include the value of this put option in their measure of cost (Kane 2014).

With regard to TARP's gross benefits, a credible evaluation of the impact of TARP assistance to financial institutions remains elusive. First, it is difficult, if not impossible, to isolate the effects of TARP from other initiatives of the Federal Reserve, Federal Deposit Insurance Corporation, and other financial regulators, or from other influences on the economy unrelated to government programs. For example, on October 14, 2008, the Capital Purchase Program was announced jointly with the Fed's Commercial Paper Funding Facility Program and FDIC's Temporary Liquidity Guarantee Program. Furthermore, it is hard to know to what extent the financial markets would have stabilized and the economy would have recovered in the absence of an activist government response. Some have argued that government support for financial institutions during the crisis confused and frightened market participants and was itself possibly a net negative for the economy. For example, Taylor (2010 p. 170; see also 2009) argues that the initial proposed structure of TARP was a further source of shock to markets as many people "were skeptical about how [the buying up of toxic assets] would work and government officials had difficulty explaining how it would work" (p. 171), but he concludes by conceding that after it became clear that TARP would take the form of capital injections, "conditions began to improve" (p. 172). Others point out that the failure of Lehman affected markets primarily by changing perceptions of the scale of loss associated with exposures to subprime and Alt-A mortgages. Lehman's derivatives were liquidated in an orderly fashion, and no major intermediary actually failed as the result of interconnections with Lehman. From that perspective, Secretary Paulson's view that the economy was teetering at the edge of Armageddon may have been a gross exaggeration.

Finally, it is possible to argue that there were additional social costs associated with the way TARP was administered and that alternative policies might have produced greater gross benefits. These questions are the topics of the next two sections.

Was TARP Administered Properly?

Corruption is a social cost, as it entails both a misallocation of resources and a diminution of justice. Did TARP adhere to objective eligibility requirements and a credibly fair and impartial process of allocation funds, or did it also reflect political influences that were unrelated to objective criteria?

The Capital Purchase Program was the first and primary initiative under TARP through which the Treasury made preferred stock purchases in qualified financial institutions. The final decision to make CPP investments rested with the Treasury, but federal banking regulators also played an important and influential role in the CPP application and approval process. The approval process began with the interested financial institution consulting with its primary federal bank regulator about being included in the CPP. The regulator assessed the applicant's strength and viability based on bank examination ratings, financial performance ratios, and other factors.[12] Institutions that were deemed to be the strongest, received presumptive approval and their application was forwarded to the Treasury's Investment Committee. Institutions deemed to be less strong required further review and were referred to the CPP council, which was comprised of representatives from the four primary banking regulators with Treasury officials as observers. Following the CPP council's evaluation, institutions that were approved by a majority of the council members were recommended to the Treasury's Investment Committee.[13] The institutions with the lowest banking ratings and poor financial ratios were deemed ineligible for participation in the CPP, received a presumptive denial recommendation, and were not forwarded to the Investment Committee.

The Office of Financial Stability reviewed documentation of applications recommended by the regulators or the CPP Council and at times collected additional information about the applicants before submitting the applications to the Investment Committee. The Investment Committee made recommendations to the Assistant Secretary for Financial Stability for final approval after completing its review (US GAO 2010). Clearly, discretionary judgments played a significant role in the approval process.[14]

[12] Six performance ratios were identified to evaluate applicants. Three related to regulatory capital levels: the Tier 1 risk-based capital ratio, total risk-based capital ratio, and Tier 1 leverage ratio. The quality of assets was assessed using the ratio of classified assets, nonperforming loans, and construction and development loans to capital and reserves.

[13] The Treasury provided guidance to the Capital Purchase Program council to use in assessing applicants that allowed consideration of additional factors (such as signed merger agreements, confirmed investments of private capital beyond, and others) beyond examination ratings and financial ratios (US GAO 2010, pp. 11–147).

[14] The nine largest financial institutions that were included in the Capital Purchase Program at the time of its establishment did not follow the application process described above. These were Bank of America, Bank of New York Mellon, Citigroup, Goldman Sachs, JP Morgan Chase, Merrill Lynch, Morgan Stanley, State Street, and Wells Fargo. They were offered assistance by virtue of their systemic importance and were asked to participate in the program even if they did not want to do so.

The US Government Accountability Office's (2010) review of the approval process for participation in the Capital Purchase Program revealed that almost all of the reviewed institutions had satisfactory or better overall ratings. However, a quarter of the examination ratings used for making approvals were more than one year old, 5 percent were more than 16 months old, and 104 of 567 reviewed applications lacked a date of the most recent bank examination. Several approved institutions also exhibited weaker characteristics that made their viability doubtful. The Government Accountability Office discovered that 12 percent of the approved cases reviewed (66 institutions) either: 1) did not meet the performance ratio guidelines; 2) had an unsatisfactory bank examination rating; or 3) had a formal regulatory enforcement action involving safety and soundness concerns. This could partly be a result of limited communication and guidance from the Treasury to the CPP council regarding how to assess viability during the early stages of the CPP. A 2009 audit of the CPP review and approval process by the Federal Reserve's Inspector General found that applicants would have been analyzed consistently and completely if the Treasury had provided formal and detailed procedures to evaluate applicants (Board of Governors 2009).

Marginal cases that were approved for the Capital Purchase Program displayed more financial weaknesses than others. The US Government Accountability Office (2010) reports that 39 percent of the 66 approved institutions with marginal characteristics missed at least one CPP dividend payment. In comparison, only 20 percent of all CPP participants had missed at least one dividend payment. By August 2010, several marginal cases also had received formal enforcement actions.

Not all of the administrative shortcomings of TARP can be attributed to innocent oversights or incompetence, and political connections seem to have played a part in the approval and allocation of TARP funds.[15] Congressional campaign contributions from the financial services industry were associated with a higher likelihood of voting in favor of the Emergency Economic Stabilization Act of 2008 (Mian, Sufi, and Trebbi 2010). Institutions that employed ex-regulators or federal government employees, or were headquartered in the election districts of House members on key finance committees were more likely to be approved for participation in the Capital Purchase Program (Duchin and Sosyura 2012; Blau, Brough, and Thomas 2013). For example, Duchin and Sosyura (2012) report that banks employing a director who worked at the Treasury or one of the banking regulators were 9.1 percentage points more likely to be approved for participation in CPP. Campaign contributions and lobbying expenditures by institutions increased the likelihood of receiving CPP investments. Political connections also influenced the amount and timing of investments under TARP. Politically connected institutions received a greater amount

[15] Some readers will remember the infamous Keating Five, a previous example where it appeared that there had been political interference in financial regulation. Five US Senators were accused of improperly intervening in 1987 on behalf of Charles H. Keating, Jr., Chairman of the Lincoln Savings and Loan Association. Lincoln was a target of regulatory investigation by the Federal Home Loan Bank Board (FHLBB). Following the intervention of the Senators, FHLBB backed off from taking action against Lincoln and subsequently it failed in 1989 at a cost of $3 billion to the taxpayers.

of TARP support, and it was provided earlier, relative to firms that lacked political connections. Politically connected recipients subsequently underperformed unconnected firms based on both stock returns and on accounting-based performance measures (Duchin and Sosyura 2012).

Alternative Policies, Inefficiencies, and Political Constraints

TARP was crafted in a volatile political and economic environment, in the middle of a financial crisis, and just prior to a major election (Swagel 2009). Its architects were in a hurry to enact TARP and knew that it was not going to be easy to get agreement on a blank check for hundreds of billions of dollars to assist "fat cats" on Wall Street. TARP's main design challenge was to balance the often conflicting objectives of shoring up banks while ensuring "social justice" by limiting how much banks' owners, creditors, and employees would benefit personally at taxpayers' expense. Here we consider several of the alleged shortcomings of TARP's design that gave rise to inefficiencies relative to alternatives, and also consider the extent to which those shortcomings were the product of political compromise.[16]

Should the Structure of TARP Have Been Debated More Broadly?

One of us suggested to a senior Congressional staff member in September 2008 that Congress should invite economists to offer views on how TARP might be structured. This could have been accomplished very quickly, as many knowledgeable people were interested in participating. The staffer explained that an election was coming. Democrats anticipated control of both houses of Congress and the White House. They had little to gain, and much to lose, from becoming vocal proponents of a new plan or vocal opponents of Secretary Paulson's plan. Although the Democratic leadership had serious doubts about the asset purchase plan, they did not want independent testimony to put them "on the spot." They did not want to have to create or politically "own" new ideas about assisting banks. The path of least political resistance was to let Secretary Paulson take the lead and the responsibility. This explains why no independent testimony or substantive public policy debate over the structure of TARP occurred during the crucial days from mid-September until early October 2008. It may also explain the Treasury's ill-fated advocacy of the asset purchase approach—an idea that was untested and viewed by many as unworkable. In contrast, capital injections had been used successfully in the United States in the 1930s and in Scandinavia in the 1990s. Problems in Japan's implementation

[16] We consider broad design features below. There are also several narrower design issues that have been considered in the literature. For example, Wilson (2013) finds that permitting some banks to issue noncumulative preferred stock was associated with a greater probability of missing a dividend payment.

of capital injections were also well known (Calomiris 1998; Calomiris and Mason 2004; Hoshi and Kashyap 2010).[17]

Those experiences provide evidence in favor of the efficacy of capital injections, and identify some design errors in TARP's capital injection program that might have been corrected. Specifically, we consider: 1) the requirement that warrants be issued alongside preferred stock, 2) permitting common dividends to be maintained by recipients of TARP assistance, 3) debt overhang problems (which ultimately led to the government's common stock holdings in Citigroup and AIG), and 4) compensation limits for recipients of assistance.

Should Warrants Have Been Required?

Requiring recipients of TARP assistance to issue warrants alongside preferred stock had political appeal as it allowed taxpayers to participate in the upside once the crisis ended. But did the use of warrants make economic sense as part of TARP assistance? The purpose of TARP was not to create profit opportunities for taxpayers, but to stabilize the banking system and the economy. From that perspective, requiring warrants was not helpful because the inclusion of warrants discouraged private stock issuance by taking away some of the upside available to stockholders (Calomiris 1998, 2009a, b; Calomiris and Mason 2004). A much better approach would have been to reward banks that received preferred stock assistance for raising new common stock in the market (for example, by making coupons on preferred stock fall with new common stock issues). That approach would have magnified the effects of TARP preferred stock through higher common stock offerings, resulting in greater bank stability and more protection against loss to taxpayers. It would have meant an even larger subsidy on the preferred stock coupon, but subsidy is the essence of government assistance—that subsidy would have been directly linked to the economic improvements that were the goal of TARP. Warrants were a popular tool for politicians who wanted to make speeches about how bankers' profiteering would be limited, but they also were an impediment to encouraging the more rapid private recapitalization of banks, which would have reduced taxpayers' risks and increased banks' stability and lending capacity.

Should Common Stock Dividends of TARP Recipients Have Been Reduced?

Participants in the Capital Purchase Program should not have been permitted to pay common stock dividends. If banks are undercapitalized enough to warrant taxpayer-funded recapitalization, then they should be forced to accumulate capital through retained earnings. Also, the protection taxpayers enjoy through the seniority of preferred shares is lessened, and debt overhang problems are exacerbated, by paying dividends.

This feature of TARP is generally explained as the result of a political deal between the Treasury and the healthy large banks (such as JP Morgan Chase) which

[17] For a summary of some of the literature on crisis-management policies, see Calomiris, Klingebiel, and Laeven (2005).

otherwise would not have bent to Treasury's pressure to participate in TARP. But that explanation raises a deeper question: what was the presumed advantage from getting healthy banks to participate in TARP? One explanation is the desire to mask differences among banks so that weak banks are not identified by virtue of their participation. But the market was well aware of the differences in the relative strength of various financial institutions. The 90-day moving average of Citigroup's market equity-to-asset ratio fell to about 2 percent in late 2008 and reached 1 percent in early 2009, while JP Morgan Chase's market equity-to-asset ratio consistently remained several times as high (Calomiris and Herring 2013). Having JP Morgan Chase sign up for assistance did nothing to make Citigroup seem stronger.

Should Compensation Limits Have Been Less Onerous?

Limits on participating banks' compensation rules were part of TARP from the beginning and the limits became more binding with the passage of ARRA in February 2009. Like the use of warrants, compensation limits served the political purpose of building support for TARP assistance programs, but increasingly binding limits encouraged strong banks to avoid TARP. That policy generated the early exodus from TARP by many big banks in mid-2009 and reduced other relatively strong banks' willingness to apply for assistance in the program (Bayazitova and Shivdasani 2012; Cadman, Carter, and Lynch 2012), which lessened the impact of TARP in increasing the supply of lending. Cadman, Carter, and Lynch (2012) find that increasing compensation from the 25th to the 75th percentile of banks was associated with a doubling of a bank's unwillingness to accept TARP funds. They also find that TARP recipients tended to suffer larger managerial turnover and the presence of severance agreements made banks hesitant to participate in TARP, consistent with concerns about a talent drain related to compensation limits.[18] Bayazitova and Shivdasani (2012, p. 390) find that the presence of highly compensated CEOs reduced the chance of being approved for TARP: "A one-standard-deviation increase in the log of CEO compensation in excess of $500,000 is associated with an 11.4-percentage point reduction in Treasury approval, or roughly one-sixth of the size of the unconditional approval probability."

Better Ways of Addressing Debt Overhang?

The debt overhang problem arises when debts are so large that any gains to banks are likely to benefit only debtholders rather than shareholders. In the cases of AIG and Citigroup, the debt overhang problem ultimately led to the transformation of government assistance into common stock ownership. Might better alternative solutions have avoided such a high degree of taxpayer exposure to potential loss? At least three viable alternatives were known and discussed. The problem with each of

[18] Cadman, Carter, and Lynch (2012) do not find any difference in lending between TARP recipients and other banks, but as they recognize, this likely reflects selectivity bias; TARP recipients likely would have cut lending if they had not received TARP. Li (2013) finds that TARP funding did in fact increase the supply of lending.

them is that they would have required an explicit payment of a subsidy rather than the implicit payment associated with TARP's more politically palatable willingness to bear downside risk.

One approach would have used out-of-the-money guarantees to boost the value of distressed assets, thereby raising the value of banks' assets and overcoming the debt overhang. One of us proposed such an approach for especially weak banks in late 2008 and early 2009 (Calomiris 2009b), and argued that such subsidies could be combined with requirements that banks receiving such guarantees raise common stock to bolster their resiliency and enable them to expand their lending. To be concrete, in late 2008, as the result of the collapse of market liquidity, many portfolios of subprime and Alt-A mortgages were being priced very low (in rarely observed market transactions) compared to their expected recovery values. If the government had offered a free put option on, say, Citigroup's entire portfolio of subprime and Alt-A mortgages and mortgage-backed securities (to prevent cherry picking) at 50 percent of face value, that would have substantially raised the market value of Citigroup's shares. Even if 50 percent of the mortgages underlying that portfolio had gone to foreclosure with a loss, given default, of 50 percent, the recovery value of the portfolio would have been 75 percent, implying no cash flow cost to taxpayers from providing a put option at 50 percent of face value. Of course, if this guarantee had been priced on market terms, there would have been no subsidy, and also no effect on Citigroup's stock price.

A second approach would be to attach put options to new stock offerings. The government could offer buyers of new shares a put option at, say, 30 percent below the price paid for those shares in the market. This step would raise the price of new offerings, substantially improving the ability of banks to raise common stock, and would limit taxpayers' exposure to extremely unlikely states of the world (where cumulative losses on shares exceeded 30 percent).

A third approach would be to copy Mexico's "Punto Final" program of 1999, which helped to end the Mexican banking system's financial gridlock (Calomiris, Klingebiel, and Laeven 2005; Calomiris 2009b). The Mexican government matched loan write-downs that were agreed between creditors and debtors so long as they were agreed quickly (within six months). For example, the US government could have agreed to pay 30 cents to a creditor for every dollar that the creditor decided to forgive in troubled mortgages, leaving it to the creditor to decide which mortgages to include in the subsidized write-down program. Value-maximizing creditors would have used this subsidy to write down mortgages that were close calls—those for which (absent the subsidy) foreclosure was the best strategy for the creditor, but for which a subsidy would make it worthwhile for the creditor to agree to a moderate write-down. A Punto Final approach not only would have raised bank asset and equity values, it would have improved the wealth of many mortgage holders and eliminated some of the uncertainty that plagued the housing and mortgage markets.

Despite discussions of all three approaches, including by Secretary Geithner in early 2009, political opposition to subsidizing the big banks blocked these subsidy proposals. Ricardo Caballero, a vocal proponent of using subsidized out-of-the-money guarantees of bank assets or stock offerings, complained in frustration in an article published in February 2009: "Politics require that a 'good deal for taxpayers' is added to . . . [the] . . . principles [guiding TARP], but the truth is that the best deal for taxpayers, once one considers the endogenous response of the economy, is anything that works to stabilize the financial system . . ."

Should Assistance to Banks Have Been More Generous or More Selective?

Li (2013) shows that TARP recipients increased the supply of credit they provided to the economy. Local markets in which a higher proportion of banks received TARP funds experienced improved economic conditions (Berger and Roman 2015). Croci, Hertig, and Nowak (2015) argue that more forgiving standards for TARP assistance to voluntary participants would have reduced resolution costs for the Federal Deposit Insurance Corporation, and that on net, this would have been desirable.

These analyses tend to support the view that TARP should have been more generous. However, there are some counterbalancing considerations. Financial institutions that can reasonably expect to receive assistance if they take risks that could lead to insolvency, will have a moral hazard incentive to engage in riskier behavior, which means that the costs of providing such incentives are potentially large (Duchin and Sosyura 2014). Furthermore, the ability to survive the crisis after receiving assistance sets too low a standard because it neglects the long-term social gains that come from transferring poorly performing banks to relatively efficient management. Berger and Roman (forthcoming) find that TARP funds were a source of major competitive advantage in local markets, and as such they could be used inappropriately to offset the disadvantages that come from poor management. Cornett, Li, and Tehranian (2013) found that relatively weak banks that received TARP tended not to make as much high-quality loans in response to receiving funding, or to reduce expenses as much, and were less likely to repay their funding. Bayazitova and Shivdasani (2012) found no evidence of certification gains from receiving Capital Purchase Program infusions, indicating little belief among those out in the market that government selections conveyed useful positive private information about bank quality.

With respect to large banks, counterfactual resolution costs from allowing failure are hard to gauge. It is hard to find an acquirer for a global behemoth, and liquidation is particularly costly for complex organizations with cross-border reach (which substantially complicates regulatory jurisdictional challenges). On the other hand, moral-hazard costs from predictable too-big-to-fail protection may be especially great (Black and Hazelwood 2013).

Conclusion

Six years after the passage of TARP, it remains hard to measure the total social costs of the assistance to banks provided under TARP programs. While TARP's passage was associated with significant improvements in financial markets and the health of financial institutions, from an economic perspective TARP could have been better designed to achieve more benefits at lower costs. Several of the design choices made under TARP—the lack of strict limits on common dividend payments, the use of strict limits on executive compensation by participants, the contingent use of common stock investments to replace preferred stock investments in especially weak, too-big-to-fail banks instead of subsidized guarantees for troubled assets or new stock issues—all reflected fundamental political obstacles that constrained the mechanisms that were chosen.

Any evaluation of TARP must look beyond its effects on GDP and recognize that democracies also value justice, which further complicates any evaluation of TARP's design. Beyond its economic costs and benefits, TARP clearly entailed other social costs. Many found assistance to bankers unjust, or insisted on attaching conditions to that assistance that weakened its effectiveness. Evidence of corruption in choosing which banks received TARP funds also added to the noneconomic social cost.

The implementation of TARP was hasty and heavily influenced by the immediate political backlash produced by the financial crisis, especially in the crucial weeks between Lehman's failure and the election. From that perspective, perhaps the clearest lesson from TARP is that it would be useful to evaluate TARP and reach agreement within our democracy about the difficult tradeoffs involved in designing crisis assistance to banks *before* another crisis is upon us. That way, our discussion of the myriad economic and noneconomic costs and benefits can be more complete, informed, and thoughtful. This is particularly important in light of the new limits that the Dodd–Frank Act of 2010 has placed on Federal Reserve assistance to troubled financial institutions under Section 13(3) of the amended Federal Reserve Act. The Fed was actively involved throughout the financial crisis in taking on risk through guarantees, purchases, and loans. In the future, the ability of the Fed to do so will be substantially more constrained. Although it is reasonable and appropriate to limit Fed discretion on fiscal matters, having done so, it is all the more necessary to plan ahead transparently and wisely for the next crisis. The United States has suffered 17 major banking crises since 1792; it is unlikely that the subprime mortgage crisis will be our last.

■ *The authors would like to thank David Autor, Allen Berger, Jerry Caprio, Stijn Claessens, Richard Herring, Chang-Tai Hseih, Edward Kane, Luc Laeven, Allan Meltzer, Phillip Swagel, and Timothy Taylor for thoughtful comments and suggestions on earlier drafts of this paper.*

References

Adrian, Tobias, and Hyun Song Shin. 2009. "Financial Intermediaries, Financial Stability and Monetary Policy." In *Maintaining Stability in a Changing Financial System*, A symposium sponsored by the Federal Reserve Bank of Kansas City, Jackson Hole, Wyoming, August 21–23, 2008, pp. 287–334.

Anari, Ali, James Kolari, and Joseph Mason. 2005. "Bank Asset Liquidation and the Propagation of the U.S. Great Depression." *Journal of Money, Credit and Banking* 37(4): 753–73.

Bayazitova, Dinara, and Anil Shivdasani. 2012. "Assessing TARP." *Review of Financial Studies* 25(2): 377–407.

Berger, Allen N., and Raluca A. Roman. 2015. "Did Saving Wall Street Really Save Main Street? The Real Effects of TARP on Local Economic Conditions." Available at SSRN: http://ssrn.com/abstract=2442070.

Berger, Allen N., and Raluca A. Roman. Forthcoming. "Did TARP Banks Get Competitive Advantages?" *Journal of Financial and Quantitative Analysis.*

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

Bernanke, Ben S., and Mark Gertler. 1989. "Banking and Macroeconomic Equilibrium." In *New Approaches to Monetary Economics*, edited by William A. Barnett and Kenneth J. Singleton, pp. 89–114. Cambridge University Press.

Bernanke, Ben S., and Cara S. Lown. 1991. "The Credit Crunch." *Brookings Papers on Economic Activity* no. 2, pp. 205–247.

Black, Lamont K., and Lieu N. Hazelwood. 2013. "The Effect of TARP on Bank Risk-Taking." *Journal of Financial Stability* 9(4): 790–803.

Blau, Benjamin M., Tyler Brough, and Diana W. Thomas. 2013. "Corporate Lobbying, Political Connections, and the Bailout of Banks." *Journal of Banking and Finance* 37(8): 3007–3117.

Board of Governors of the Federal Reserve System. 2009. "Audit of the Board's Processing of Applications for the Capital Purchase Program under the Troubled Asset Relief Program." Board of Governors, Office of Inspector General Washington, DC, September 30.

Braithwaite, Tom, and Francesco Guerrera. 2010. "U.S. Treasury Sells Remaining Citi Shares." *Financial Times*, December 7.

Brunnermeier, Markus K., and Lasse Heje Pedersen. 2009. "Market Liquidity and Funding Liquidity." *Review of Financial Studies* 22(6): 2201–38.

Caballero, Ricardo J. 2009. "An Insurance Complement to TARP II." *Wall Street Journal*, February 17.

Cadman, Brian, Mary Ellen Carter, and Luann J. Lynch. 2012. "Executive Compensation Restrictions: Do They Restrict Firms' Willingness to Participate in TARP? *Journal of Business Finance and Accounting* 39(7–8): 997–1027.

Calomiris, Charles W. 1993. "Financial Factors and the Great Depression." *Journal of Economic Perspectives* 7(2): 61–85.

Calomiris, Charles W. 1998. "Revitalizing Ailing Japanese Banks." *Nikko Capital Trends*, May.

Calomiris, Charles W. 2007. "Devaluation with Contract Redenomination in Argentina." *Annals of Finance* 3(1): 155–92.

Calomiris, Charles W. 2008. "A Matched Preferred Stock Plan for Government Assistance." *Financial Times* Economists' Forum, September 19.

Calomiris, Charles W. 2009a. "The Subprime Turmoil: What's Old, What's New, and What's Next." In *Maintaining Stability in a Changing Financial System,* A symposium sponsored by the Federal Reserve Bank of Kansas City, Jackson Hole, Wyoming, August 21–23, 2008, pp. 19–110.

Calomiris, Charles W. 2009b. "Helping Wall Street—And Main Street." *Forbes.com*, January 21.

Calomiris, Charles W., Robert A. Eisenbeis, and Robert E. Litan. 2011. "Financial Crisis in the US and Beyond." In *The World in Crisis: Insights from Six Shadow Financial Regulatory Committees from Around the World*, edited by Robert Litan, 1–60. Wharton Financial Institutions Center.

Calomiris, Charles W., and Stephen H. Haber. 2014. *Fragile by Design: The Political Origins of Banking Crises and Scarce Credit.* Princeton University Press.

Calomiris, Charles W., and Richard J. Herring. 2013. "How to Design a Contingent Convertible Debt Requirement That Helps Solve Our Too-Big-to-Fail Problem." *Journal of Applied Corporate Finance* 25(2): 36–62.

Calomiris, Charles W., and Charles M. Kahn. 1991. "The Role of Demandable Debt in Structuring Optimal Banking Arrangements." *American Economic Review* 81(3): 497–513.

Calomiris, Charles W., Daniela Klingebiel, and Luc Laeven. 2005. "Financial Crisis Policies and Resolution Mechanisms: A Taxonomy from Cross-Country Experience." In *Systemic Financial Crises: Containment and Resolution,* edited by Patrick Honohan and Luc Laeven, 25–75. Cambridge University Press.

Calomiris, Charles W., and Joseph R. Mason.

2003. "Consequences of Bank Distress during the Depression." *American Economic Review* 93(3): 937–47.

Calomiris, Charles W., and Jason R. Mason. 2004. "How to Restructure Failed Banking Systems: Lessons from the United States in the 1930s and Japan in the 1990s." Chap. 14 in *Governance, Regulation, and Privatization in the Asia-Pacific Region*, edited by Takatoshi Ito and Anne O. Krueger, 375–420. University of Chicago.

Calomiris, Charles W., Joseph R. Mason, Marc Weidenmier, and Katherine Bobroff. 2013. "The Effects of Reconstruction Finance Corporation Assistance on Michigan Banks' Survival in the 1930s." *Explorations in Economic History* 50(4): 526–47.

Congressional Budget Office. 2009. "The Troubled Asset Relief Program: Report on Transactions through December 31, 2008." Washington, DC.

Cornett, Marcia Millon, Lei Li, and Hassan Tehranian. 2013. "The Performance of Banks around the Receipt and Repayment of TARP Funds: Over-achievers versus Under-achievers." *Journal of Banking and Finance* 37(3): 730–46.

Covitz, Daniel, Nellie Liang, and Gustavo A. Suarez. 2013. "The Evolution of a Financial Crisis: Collapse of the Asset-Backed Commercial Paper Market." *Journal of Finance* 68(3): 815–48.

Croci, Ettore, Gerard Hertig, and Eric Nowak. 2015. "Decision Making during the Crisis: Why Did the Treasury Let Commercial Banks Fail?" European Corporate Governance Institute (ECGI)-Law Working Paper No. 281/2015, January. Available at SSRN: http://ssrn.com/abstract=2557717.

Diamond, Douglas W. 1984. "Financial Intermediation and Delegated Monitoring." *Review of Economic Studies* 51(3): 393–414.

Diamond, Douglas W., and Raghuram G. Rajan. 2009. "The Credit Crisis: Conjectures about Causes and Remedies." *American Economic Review* 99(2): 606–10.

Duchin, Ran, and Denis Sosyura. 2012. "The Politics of Government Investment." *Journal of Financial Economics* 106(1): 24–48.

Duchin, Ran, and Denis Sosyura. 2014. "Safer Ratios, Riskier Portfolios: Banks' Response to Government Aid." *Journal of Financial Economics* 113: 1–28.

Goldman Sachs. 2008. "Berkshire Hathaway to Invest $5 Billion in Goldman Sachs." Press Release, September 23.

Gorton, Gary B. 2009. "The Panic of 2007." In *Maintaining Stability in a Changing Financial System,* A Symposium Sponsored by the Federal Reserve Bank of Kansas City, Jackson Hole, Wyoming, August 21–23, 2008, pp. 133–262.

Gorton, Gary, and Andrew Metrick. 2012. "Securitized Banking and the Run on Repo." *Journal of Financial Economics* 104(3): 425–51.

Griffen, Donal. 2011. "Citigroup Warrants Sold by Treasury for $312 Million." BloombergBusiness, January 26.

Heider, Florian, Marie Hoerova, and Cornelia Holthausen. Forthcoming. "Liquidity Hoarding and Interbank Market Spreads: The Role of Counterparty Risk." *Journal of Financial Economics.*

Hoshi, Takao, and Anil K Kashyap. 2010. "Will the U.S. Bank Recapitalization Succeed? Eight Lessons from Japan." *Journal of Financial Economics* 97(3): 398–417.

Hovakimian, Armen, Edward J. Kane, and Luc Laeven. 2012. "Variation in Systemic Risk at US Banks During 1974–2010." May 29. Available at SSRN: http://ssrn.com/abstract=2031798.

Ivashina, Victoria, and David Scharfstein. 2010. "Bank Lending during the Financial Crisis of 2008." *Journal of Financial Economics* 97(3): 319–38.

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): 305–360.

Kim, Dong H., and Duane Stock. 2012. "The Impact of the TARP Financing Choice on Existing Preferred Stock." *Journal of Corporate Finance.* 18(5): 1121–42.

Kane, Edward J. 2014. "Hair of the Dog That Bit Us: The Insufficiency of New and Improved Capital Requirements." Available at SSRN: http://ssrn.com/abstract=2225432.

Krasa, Stefan, and Anne P. Villamil. 1992. "Monitoring the Monitor: An Incentive Structure for a Financial Intermediary." *Journal of Economic Theory* 57(1): 197–221.

Kroszner, Randall. 1999. "Is It Better to Forgive Than to Receive? Repudiation of Gold Indexation Clause in Long-Term Debt during the Great Depression." Working paper, Booth School, University of Chicago, November.

Li, Lei. 2013. "TARP Funds Distribution and Bank Loan Supply." *Journal of Banking and Finance* 37(12): 4777–92.

Mason, Joseph R. 2001. "Reconstruction Finance Corporation Assistance to Financial Intermediaries and Commercial and Industrial Enterprises in the United States, 1932–1937." In *Resolution of Financial Distress: An International Perspective on the Design of Bankruptcy Laws,* edited by Stijn Claessens, Simeon Djankov, and Ashoka Mody, 167–204. World Bank Group.

Mian, Atif, Amir Sufi, and Francesco Trebbi. 2010. "The Political Economy of the US Mortgage Default Crisis." *American Economic Review* 100(5): 1967–98.

Myers, Stewart C. 1977. "Determinants of Corporate Borrowing." *Journal of Financial Economics* 5(2): 147–75.

Ng, Jeffrey, Florin P. Vasvari, and Regina Wittenberg-Moerman. Forthcoming. "Media Coverage and the Stock Market Valuation of TARP Participating Banks." *European Accounting Review.*

Philippon, Thomas, and Philipp Schnabl. 2013. "Efficient Recapitalization." *Journal of Finance* 68(1): 1–42.

Schumer, Charles. 2008. "How to Rescue the Banks." *Wall Street Journal*, October 14.

Schwarz, Krista. 2015. "Mind the Gap: Disentangling Credit and Liquidity in Risk Spreads." Working paper, University of Pennsylvania Wharton School.

Swagel, Phillip. 2009. "The Financial Crisis: An Inside View." *Brookings Papers on Economic Activity*, Spring.

Taliaferro, Ryan. 2009. "How Do Banks Use Bailout Money? Optimal Capital Structure, New Equity and the TARP." Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1481256.

Taylor, John B. 2009. *Getting Off Track: How Government Actions and Interventions Caused, Prolonged, and Worsened the Financial Crisis.* Hoover Press.

Taylor, John. B. 2010. "Getting Back on Track: Macroeconomic Policy Lessons from the Financial Crisis." Federal Reserve Bank of St. Louis *Review* (May/June): 165–76.

US Department of the Treasury. Nd. "Treasury Prices Sale of Citigroup Subordinated Notes for Proceeds of $894 Million." Last updated, 5/30/2013. http://www.treasury.gov/initiatives/financial-stability/news-room/news/Pages/TREASURY-PRICES-SALE-OF-CITIGROUP-SUBORDINATED-NOTES-FOR-PROCEEDS-OF-$894-MILLION.aspx.

US General Accountability Office (GAO). 2009. *Troubled Asset Relief Program: One Year Later, Actions Are Needed to Address Remaining Transparency and Accountability Challenges.* GAO-10-16. Washington, DC.

US General Accountability Office (GAO). 2010. "Troubled Asset Relief Program: Opportunities Exist to Apply Lessons Learned from the Capital Purchase Program to Similarly Designed Programs and to Improve the Repayment Process." GAO-11-47. Washington, DC.

US General Accountability Office (GAO). 2011a. *Troubled Asset Relief Program: Status of Programs and Implementation of GAO Recommendations.* GAO-11-74. Washington, DC.

US General Accountability Office (GAO). 2011b. "Troubled Asset Relief Program: Status of GAO Recommendations to Treasury." GAO-11-906R, September 16. Washington, DC.

US General Accountability Office (GAO). 2012a. "Capital Purchase Program: Revenues Have Exceeded Investments, But Concerns about Outstanding Investments Remain." GAO-12-301. Washington, DC.

US General Accountability Office (GAO). 2012b. "Troubled Asset Relief Program: Government's Exposure to AIG Lessens as Equity Investments Are Sold." GAO-12-574. Washington, DC.

US General Accountability Office (GAO). 2013a. "Troubled Asset Relief Program: Treasury Sees Some Returns as It Exits Programs and Continues to Fund Mortgage Programs." Washington, DC. GAO-13-192.

US General Accountability Office (GAO). 2013b. "Troubled Asset Relief Program: Status of GAO Recommendations to Treasury." GAO-13-324R. March 8. Washington, DC.

US General Accountability Office (GAO). 2014. "Troubled Asset Relief Program: Status of the Wind Down of the Capital Purchase Program." GAO-14-388. Washington, DC.

Veronesi, Pietro, and Luigi Zingales. 2010. "Paulson's Gift." *Journal of Financial Economics* 97(3): 339–68.

Wilson, Linus. 2013. "TARP's Deadbeat Banks." *Review of Quantitative Finance and Accounting* 41(4): 651–74.

Journal of Economic Perspectives—Volume 29, Number 2—Spring 2015—Pages 81–106

AIG in Hindsight[†]

Robert McDonald and Anna Paulson

The near-failure on September 16, 2008, of American International Group (AIG) was an iconic moment of the financial crisis. AIG, a global insurance and financial company with $1 trillion in assets, lost $99.3 billion during 2008 (AIG 2008, p. 194) and was rescued with the help of the Federal Reserve, the Federal Reserve Bank of New York, and the US Treasury. The rescue played out over many months and involved the extension of loans, the creation of special purpose vehicles, and equity investments by the Treasury, with the government assistance available to AIG ultimately totaling $182.3 billion. The decision to rescue AIG was controversial at the time and remains so. AIG's fate also provided an important touchstone in discussions of financial reform. AIG motivated the enactment of new rules governing nonbank financial institutions, as well as rules about the treatment of financial derivatives.

In this paper, we begin with an overview of AIG's main corporate financial indicators from 2006–2009. However, most of the attention paid to AIG—and our focus—concerns the two main activities that caused the insurance company to be driven to the edge of bankruptcy by falling real estate prices and mortgage foreclosures: AIG's securities lending business and its credit default swap business. Although much of the discussion concerning AIG has centered on its credit default swap business, we will show that losses from its securities lending business

■ *Robert McDonald is the Erwin P. Nemmers Professor of Finance, Kellogg School of Management, Northwestern University, Evanston, Illinois. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Anna Paulson is Vice President and Director of Financial Research, Federal Reserve Bank of Chicago, Chicago, Illinois. Their email addresses are r-mcdonald@northwestern.edu and anna.paulson@chi.frb.org.*

[†] To access the Appendix, Data Appendix, and disclosure statements, visit
http://dx.doi.org/10.1257/jep.29.2.81

doi=10.1257/jep.29.2.81

were of a similar magnitude. On September 16, 2008, the cumulative losses from these two activities were on the order of $50 billion, and both appear to have played important roles in AIG's near-failure (as also emphasized by Pierce 2014; Taibbi 2011, chap. 3).

We then turn to a description of the government rescue of AIG, including the special purpose vehicles "Maiden Lane II" and "Maiden Lane III" that the New York Fed created to deal with the assets related to AIG's securities lending and credit default swap operations, respectively. In particular, we examine the write-downs on the assets in these portfolios from each asset's inception to October 2014. AIG's real estate positions were apparently motivated by the belief that these investments would not default. The analysis sheds light on a claim often made by AIG executives that their mortgage-related investments might have suffered a decline in their market value in the short-term, but that they would pay off over time. This claim implicitly attributes any price decline in such securities to short-term illiquidity. The head of the AIG Financial Products subsidiary, Joseph Cassano, often referred to the mortgage-related securities that AIG insured through credit default swaps as "money good" (for example, see American International Group Investor Meeting 2007). Mark Hutchings (2010), who ran AIG's securities lending business, made similar statements about the real estate–related investments financed by securities lending. However, this stark claim that assets were "money good" is not borne out: a number of AIG's mortgage-related investments suffered principal write-downs. In our concluding section, we discuss the question of how to think about AIG as a financial firm.

It is important to be clear about what we do not do in this paper. We do not analyze AIG's regulatory oversight prior to the crisis. We discuss what happened in the AIG rescue, but we do not analyze alternative policies or capital structures for a rescue. We discuss the specific parties who benefited most from the rescue, but we do not address the broad question of what might have happened to the financial system had AIG failed. There was certainly reason for concern: In testimony about the AIG rescue, Federal Reserve Chairman Ben Bernanke noted that AIG had $20 billion of commercial paper outstanding and $50 billion of exposure to other banks via loans, lines of credit, and derivatives. Lehman Brothers had around $5.7 billion in commercial paper, and its failure wreaked havoc on money market mutual funds (FDIC 2011). Policymakers and academics have written extensively about potential systemic consequences from the failure of a large, interconnected financial firm like AIG: for example, Acharya, Gale, and Yorulmazer (2011), Brunnermeier and Pedersen (2009), Kacperczyk and Schnabl (2010), Duarte and Eisenbach (2014), and Ellul, Jotikasthira, Lundblad, and Wang (2014), among many others.

AIG Financials: 2006–2009

AIG was an international insurance conglomerate with four main lines of business: 1) General Insurance, including property/casualty and commercial/industrial

Table 1
AIG Financial Indicators by Operating Segment, 2006–2009
(billions of dollars)

Item	*2006*	*2007*	*2008*	*2009*
Revenues	113.39	110.06	11.10	96.00
Earnings	14.05	6.20	−99.29	−12.31
Realized capital gains	0.11	−3.59	−55.48	−6.86
Unrealized CDS losses (AIGFP)	0	−11.47	−28.60	1.42
Operating Income				
General Insurance	10.41	10.53	−5.75	0.17
Life Insurance & Retirement Services	10.12	8.19	−37.45	2.04
Financial Services	0.38	−9.52	−40.82	0.52
Asset Management	1.54	1.16	−9.19	NA
Assets				
General Insurance	167.00	181.71	165.95	154.73
Life Insurance & Retirement Services	550.96	613.16	489.65	553.49
Financial Services	202.49	193.98	167.06	132.82
Asset Management	78.28	77.27	46.85	NA

Sources: AIG 2008 10-K, pp. 71, 194, and 225 and AIG 2009 10-K, pp. 72, 195, and 230.
Notes: In 2009, results from asset management activities were included in the Life Insurance & Retirement Services category. Revenue is composed of premiums and other income, net investment income, realized capital gains (or losses), and unrealized credit default swap (CDS) losses. Earnings are equal to net income (or losses) as reported on AIG's consolidated statement of income. Realized capital gains are primarily comprised of sales of securities and other investments, foreign exchange transactions, changes in the fair value of non-AIGFP derivative instruments that do not qualify for hedge accounting treatment, and other-than-temporary impairments on securities. Unrealized CDS losses are the unrealized market valuation loss on AIGFP's super senior credit default swap portfolio. Operating income is equal to pre-tax income (or loss) for each business segment. Assets are equal to year-end identifiable assets for each business segment.

insurance; 2) Life Insurance and Retirement, including individual and group life insurance and annuities; 3) Asset Management, including private banking, brokerage, and investment advisory services; and 4) Financial Services, including a capital markets division, consumer finance, and aircraft leasing. Looking at that list of lines of business, it is not at all obvious why AIG had significant exposure to risks from falling real estate prices and default rates on subprime mortgages.

Each year, public firms must file a 10K report with the Securities and Exchange Commission with an in-depth presentation of their financial position. In its 2007 10K report, AIG listed $1.06 trillion in assets (AIG 2007b, p. 130). Table 1 presents financial indicators for 2006–09, which help to put AIG's 2008 performance into perspective. The firm was showing some reasons for concern in 2007, including losses in the Financial Services division and unrealized losses in its credit default swap business. But in 2008, AIG lost money in all of its main lines of business, with the largest losses in the Life Insurance and Financial Services divisions. In both cases, the losses stemmed from heavy bets on real estate–related financial products.

The Life Insurance division lost money primarily because of securities lending ($21 billion in losses), where life insurance company assets were loaned in exchange for cash that was used to invest in mortgage-related securities. In the case of financial services, AIG had written credit default swaps on mortgage-related bonds, losing $28.6 billion in 2008 (AIG 2008, p. 265). The securities lending business will be discussed in the next section; the credit default swap business will be discussed in the section after that. AIG's reported 2008 revenue of $11.1 billion incorporates the losses from securities lending, credit default swaps, and other sources.

AIG's Securities Lending Business

During 2008, AIG's life insurance subsidiaries lost approximately $21 billion from securities lending, in which the life insurance subsidiaries loaned out assets and invested the proceeds in risky assets, including assets backed by subprime residential mortgage loans. In this section, we discuss AIG's securities lending activity, which created unique problems because of its links to AIG's state-regulated life insurance subsidiaries. Recently, Pierce (2014) has examined the securities lending business in detail. We argue that it is impossible to evaluate the potential consequences of an AIG failure without understanding AIG's life insurance and securities lending activities.

What Is Securities Lending?

In a securities lending transaction, one party borrows a security from another and deposits collateral, typically cash, with the securities lender. The borrower may use the security as part of a short-selling strategy or to deliver a particular security to a customer. The securities lender invests the cash collateral and earns a yield from these investments, less a rebate paid to the securities borrower. Absent default, the lender remains the economic owner of the security that is on loan, earning its return including any dividend or coupon payments. The cost to the security borrower is the difference between the return the borrower could have earned investing the cash collateral and the rebate fee, which is a market price determined by the scarcity of the security on loan. The term of a securities lending transaction may extend for various periods up to several months, but in many cases either party can terminate the transaction early. The borrower can end the transaction by returning the security to the lender, at which time the lender must also return the cash deposit to the borrower. A problem can arise if many borrowers simultaneously decide to end transactions and the securities lender does not have, or cannot raise, sufficient cash to meet these demands in a timely fashion.[1]

[1] Securities lending transactions are very similar to repurchase agreements, as discussed in Adrian, Begalle, Copeland, and Martin (2013). For additional background on securities lending, see Aggarwal, Saffi, and Sturgess (2012) and Bank of England (2010).

Characteristics of AIG's Securities Lending

AIG's securities lending activities were conducted "primarily for the benefit of certain AIG insurance companies" (AIG 2007b, p. 108). These activities were centralized in a noninsurance subsidiary, AIG Global Securities Lending (GSL), which served as an agent for AIG's subsidiary life insurance companies. The life insurance companies provided securities, primarily corporate bonds, to GSL. These securities were loaned to banks and broker-dealers in return for cash collateral that was invested by GSL. The investment proceeds were used to fund the rebate to the security borrower, and the remainder was split 50–50 between GSL and the insurance companies. Nearly all of AIG's security loans had a one-month term (Hutchings 2010).[2]

AIG expanded its securities lending rapidly in the run-up to 2008. At the end of 2003, the firm had less than $30 billion in securities lending outstanding. At the peak in 2007Q3, AIG had securities lending outstanding of $88.4 billion (AIG 2007a, p. 2). AIG had securities lending of $70 billion during the second quarter of 2008, which then fell almost to zero by the fourth quarter of 2008.

AIG consistently lent more than 15 percent of its domestic life insurance assets: in 2007, for example, the figure was 19 percent. By comparison, Metlife, another active insurance securities lender, never had more than 10 percent of its domestic life insurance assets on loan.

Typically, securities lending collateral is invested in short-term, highly liquid securities: A firm cannot easily lend its securities for cash collateral if possible borrowers of those securities fear that their cash collateral may not be secure. However, AIG invested a substantial portion of the cash collateral it received from securities borrowers in longer-term, illiquid instruments, including securities dependent on the performance of subprime residential mortgages. At the end of 2007, 65 percent of AIG's securities lending collateral was invested in securities that were sensitive either directly or indirectly to home prices and mortgage defaults. These securities included some backed by residential and commercial mortgages, as well as others backed by credit card, auto, and home equity loans. It also included collateralized debt obligations (CDOs), which are structured financial instruments that are backed by a pool of financial assets, often the riskier tranches of mortgage-backed securities. Cash flows to collateralized debt obligations are divided into tranches ranked from junior to senior. Any losses are first allocated to the more junior tranches until their value is exhausted, a structure which offers a degree of protection to senior tranches.

Of the remainder of AIG's securities lending collateral, 19 percent was invested in corporate bonds and 16 percent was in cash or other short-term investments (AIG 2007b, p. 108). For comparison, a Risk Management Association (2007) survey of securities lenders shows that on average 33 percent of lending proceeds was invested

[2] Term arrangements can be fixed or indicative. If they are indicative, they can be terminated early without penalty (Bank of England 2010). We do not have information about whether AIG's arrangements were fixed or indicative.

in mortgage-backed securities, asset-backed securities (a broad category of securities backed by credit card receivables, auto loans, and the like), and collateralized debt obligations, with the remainder invested 42 percent in corporate bonds and 25 percent in cash and short-term investments.

AIG's use of securities lending collateral to purchase residential mortgage-backed securities and collateralized debt obligations is similar to the broader phenomenon described in Krishnamurthy, Nagel, and Orlov (2014) of financial firms using short-term funding like repurchase agreements and securities lending to fund assets that had previously been funded through insured bank deposits. AIG's investments of securities lending collateral in real estate–related instruments accelerated after 2005. On the other hand, the AIG Financial Products (AIGFP) subsidiary decided to stop increasing its exposure to real estate–related risk near the end of 2005. It took some time to implement this decision, however, and deals that were in the pipeline were completed, and as a result AIGFP's real estate exposure continued to grow. In addition, some of the collateralized debt obligations that AIGFP insured were "actively managed," which meant that the manager of the security could replace maturing, refinanced, and defaulting mortgages with new ones, including the particularly default-prone mortgages that were made in 2006 and 2007.

The AIG securities lending business was characterized by a large liquidity and maturity mismatch. Securities borrowers can demand the return of their cash collateral on short notice. However, AIG was investing this cash in long-term assets whose market values and liquidity could vary substantially in the short run. As long as AIG could make new security loans when existing ones came due, it could maintain its investments in long-run, illiquid assets. But an arrangement based on a liquidity and maturity mismatch, like this one, is clearly vulnerable to bank-run dynamics. The security borrowers have incentives that are similar to bank depositors who lack deposit insurance. Depositors will rush to withdraw cash when they are concerned about their bank's solvency. They want to make sure that they get their funds before the bank runs out of money. Similarly, security borrowers who are worried about the AIG's ability to return their cash on demand are likely to ask for it to be returned. Efforts to satisfy these demands will further erode AIG's liquidity and generate losses that will prompt other securities borrowers to demand the return of their cash collateral.

Indeed, before AIG was rescued on September 16, 2008, securities lending counterparties began to terminate these lending agreements. Standard and Poor's, Moody's, and Fitch all lowered AIG's credit rating in May or June 2008. AIG announced large second-quarter losses on August 6, 2008. The possibility of further losses and still-lower credit ratings appears to have accelerated the efforts of counterparties to reduce their securities lending exposure to AIG. Because the combination of falling real estate prices and higher mortgage foreclosures had reduced the market price of securities tied to these underlying assets, and because it did not have access to other sources of liquidity, AIG was unable to generate sufficient funds to meet redemption requests and to return the cash collateral. Moreover, its losses on securities lending threatened the regulatory capital positions of AIG's life insurance subsidiaries, a point we discuss later and one that is also emphasized by Pierce (2014).

Like many episodes during the crisis, AIG's securities lending problems can be viewed through the lenses of both liquidity and solvency. AIG (2008, p. 4) summed up its dilemma with respect to securities lending with considerable understatement in its 2008 10K report: "During September 2008, borrowers began in increasing numbers to request a return of their cash collateral. Because of the illiquidity in the market for RMBS [residential mortgage-backed securities], AIG was unable to sell RMBS at acceptable prices and was forced to find alternative sources of cash to meet these requests." On Monday, September 15, 2008, alone, AIG experienced returns under its securities lending programs that led to cash payments of $5.2 billion (AIG 2008, p. 4).

On September 16, 2008, AIG received "alternative sources of cash" from the Federal Reserve Bank of New York. The cash was initially in the form of loans. However, the New York Fed soon set up several limited liability companies as financial vehicles to handle its rescue of AIG. In December 2008, one of these companies called Maiden Lane II purchased AIG's remaining portfolio of residential mortgage-backed securities, in which it had invested securities lending collateral, for $20.5 billion—a 48 percent discount relative to their par value of $39.3 billion. According to the Congressional Oversight Panel report (2010, p. 45), AIG's securities lending counterparties demanded the return of $24 billion in cash collateral between September 12 and September 30, 2008. Ultimately, AIG reported losses from securities lending in excess of $20 billion in 2008.

Securities Lending and Bankruptcy

What would have happened to AIG's insurance companies and securities lending counterparties in the event of an AIG bankruptcy? Generally, if a securities lender seeks bankruptcy protection, the borrower simply takes ownership of the security that it borrowed; any additional claims associated with the transaction would be resolved in bankruptcy. The value of the security on loan is marked to market daily, and the collateral is adjusted accordingly, so any additional claims if a security lender goes bankrupt would typically be small. Because securities lending transactions are exempt from the "automatic stay" provisions of the bankruptcy code—that is, the rule that once bankruptcy has been declared, creditors cannot move to collect what they are owed—resolving these securities lending transactions should be fast and straightforward.

However, AIG's securities lending was conducted largely on behalf of its life insurance companies, which were regulated at the state level. If AIG had declared bankruptcy, the resolution of claims related to securities lending would likely have depended on the actions of state insurance regulators. When a life insurance company cannot meet its financial obligations, a state insurance commissioner will take control of the company's operations and place it in receivership.[3] Federal

[3] The state receivership process has three stages: 1) conservation, 2) rehabilitation, and 3) liquidation. The receivership process can involve transfers of blocks of assets and liabilities to other companies. If the company cannot be rehabilitated or sold, it is declared insolvent and the commissioner liquidates the company and distributes assets or the proceeds from asset sales to approved claimants in the manner prescribed by the state's receivership laws.

bankruptcy law does not apply to insurance companies, although the actions taken under state receivership statutes are generally patterned after federal bankruptcy. However, certain important exceptions to this practice may have been material for AIG in 2008.

If AIG had sought bankruptcy protection, state insurance commissioners would probably have seized AIG's insurance subsidiaries (Dinallo 2010). In these circumstances, the status of securities lending transactions might have varied depending on where a particular AIG insurance subsidiary was located. As of 2008, of the ten states where AIG's life insurance subsidiaries were located, only Texas had passed a version of the Insurer Receivership Model Act (IRMA) written by the National Association of Insurance Commissioners (NAIC), which allows securities lending and other qualified financial contracts to receive the same exemption from the automatic stay provisions in an insurance resolution that would apply in bankruptcy.[4] Texas-domiciled companies supplied the securities for 58 percent of AIG's securities lending. However, the legal treatment of counterparties to the remaining 42 percent of the securities supplied by life insurers located in other states would have been uncertain in an insurance insolvency. AIG's 2007 10K points out that "the securities on loan as well as all of the assets of the participating companies are generally available to satisfy the liability for collateral received" (AIG 2007b, p. 108).

An additional protection for some securities borrowers would have arisen from a unique aspect of AIG's lending program. Rather than the typical practice of requiring collateral of 102 percent of the value of the security being lent, AIG began lending securities with less than 100 percent collateral, with the AIG parent company making up the difference to the insurance subsidiary (AIG 2008, p. 3). AIG seems to have accelerated this practice as its liquidity issues grew more acute. For example, in an August 14, 2008, email, a Federal Reserve Bank of New York employee noted that "CSG [Credit Suisse Group] does not need the securities it borrows but instead AIG is using the deals to raise cash. As such CSG is looking to take a haircut on AIG's securities as opposed to posting cash to AIG in excess of the securities value which is the market standard" (available at http://fcic-static.law.stanford.edu/cdn_media/fcic-docs/2008-09-12%20FRBNY%20Email%20re%20AIG%20Meeting%20with%20OTS.pdf). By 2008, AIG had also boosted rebate fees paid to securities borrowers and was making losses on securities lending arrangements but felt this was warranted in order to avoid a "run on the bank" scenario (Hutchings 2010).

When the borrowing firm does not post enough cash to fund "substantially all of the cost of purchasing replacement assets," then from an accounting perspective, the transaction will be treated as a sale, rather than as a securities lending transaction. AIG (2008, p. 166) reported losses of $2.4 billion on securities lending transactions that had to be reclassified as "sales" in 2008.

Overall, this analysis suggests that losses for AIG's securities lending counterparties would have been small had AIG sought bankruptcy protection and if the

[4] See Fitch Ratings (2006) and "Expanding Insurance Regulation One State at a Time," available at http://www.law360.com/articles/295760/expanding-insurance-regulation-one-state-at-a-time.

counterparties were able to take possession of the securities that they had borrowed. Securities borrowers who held securities worth more than the cash they were due from AIG would not have suffered losses in an AIG bankruptcy, barring uncertainties associated with state insurance law. Note that this conclusion only takes into account the potential for direct losses. Counterparties needing to unwind or liquidate positions quickly might have suffered indirect losses as well.

Impact of Securities Lending on AIG's Domestic Life Insurance Subsidiaries

The losses for life insurance companies engaged in securities lending can be attributed to two factors: losses on sales of assets incurred when those securities were sold for cash when borrowed securities were being returned, and unrealized mark-to-market losses on similar assets that had not yet been sold. Together, these losses put AIG's domestic life insurance companies under considerable regulatory pressure. Life insurance regulators establish minimum levels of capital that take into account each company's asset risk, insurance risk, market risk, interest rate risk, and business risk (along with an adjustment to account for the fact that these risks are not perfectly correlated). When capital falls below a certain threshold, state insurance regulators are required to intervene to protect policyholders.

Looking at their official end-of-the-year balance sheets, AIG's life insurance subsidiaries appear to have made it through 2008 with a comfortable cushion of capital relative to regulatory minimums. However, these figures include over $19 billion in capital infusions in the third and fourth quarters of 2008 that were only possible because of the rescue of AIG. Table 2 shows the capital positions of the eleven AIG life insurance subsidiaries that had more than $5 billion in assets at the end of 2007. For each company, the table shows 2007 assets and the share of those assets that were on loan through AIG's securities lending business, securities lending losses in 2008, and the company's regulatory capital as of the end of 2008, both with and without the capital infusions made possible by the rescue. Eight of these eleven companies would have had negative capital without the capital infusions. The rescue funds recapitalized the life insurance companies and kept them solvent, despite their securities lending losses. This ultimately benefited AIG's life insurance policyholders.

The urgency of the problems in AIG's life insurance subsidiaries is reflected in the rapidity with which they were recapitalized: by September 30, 2008, just 14 days after the initial loan to AIG, $13.3 billion of the loan proceeds from the Federal Reserve Bank of New York had already gone toward recapitalizing the life insurance subsidiaries (Congressional Oversight Panel 2010, p. 84). Ultimately, at least $58 billion of the total government assistance to AIG went to addressing problems related to securities lending: $19 billion in capital infusions to the life insurance subsidiaries to address securities lending losses; $36.7 billion to repay collateral to securities lending counterparties ($19.5 billion from Maiden Lane II plus $17.2 billion from the revolving credit facility that the New York Fed established in the initial stages of the rescue) as well as an additional $3.1 billion from the revolving credit facility to repay securities obligations (Congressional Oversight Panel 2010, p. 237).

Table 2

The Role of the Rescue in Recapitalizing AIG's Life Insurance Subsidiaries

		2007		2008			
Company	*State*	*Assets ($ millions)*	*% of Assets in securities lending*	*Realized securities lending losses ($ millions)*	*Post-rescue capital infusions ($ millions)*	*Regulatory capital with rescue ($ millions)*	*Regulatory capital without rescue ($ millions)*
ALICO	DE	101,632	4.5%	470	967	4,332	3,365
VALIC	TX	63,999	15.1%	3,563	3,621	2,940	−681
AIG Annuity	TX	50,553	39.7%	7,109	6,048	3,242	−2,806
American General Life	TX	33,682	31.3%	3,790	3,084	2,844	−240
SunAmerica Life	AZ	39,455	27.1%	2,281	1,366	4,805	3,439
AIG SunAmerica Life	AZ	35,072	6.1%	425	281	1,317	1,036
AIG Life	DE	10,790	23.6%	870	679	465	−214
American General Life & Accident	TN	9,134	33.9%	977	786	594	−192
First SunAmerica	NY	6,479	30.3%	654	947	550	−397
American International	NY	7,093	35.1%	771	801	458	−343
United States Life	NY	5,315	25.1%	395	456	305	−151
Total: AIG Life		**364,770**	**19.0%**	**21,305**	**19,036**	**22,393**	**3,357**

Sources: Authors' calculations from insurance regulatory filings accessed through SNL Financial and March 5, 2009, Hearing before the Senate Committee on Banking, Housing, and Urban Affairs, http://www.gpo.gov/fdsys/pkg/CHRG-111shrg51303/pdf/CHRG-111shrg51303.pdf (page 43). Table includes details for active securities lending participants with assets of at least $5 billion. The "Total: AIG Life" row includes all AIG life insurance subsidiaries.

AIG's Credit Default Swap Portfolio

We now turn to AIG's credit default swap business, with the goal of understanding the position in which AIG and its counterparties found themselves on September 16, 2008.

Credit Default Swaps

A credit default swap is a derivative financial instrument that behaves like an insurance contract on a bond or a similar financial security. The writer of the credit default swap, who is the insurance seller, promises to pay to the buyer of a credit default swap the difference between the market value and the par value of the insured bond if a "credit event" occurs. For present purposes, setting aside the sometimes arcane details of these contracts, it is sufficient to think of a credit event as the failure of the bond to make a promised payment, as in a default. There are two ways that the writer of a credit default swap like AIG can suffer a loss. Obviously, a loss can occur if a credit event means that the bond or security no longer makes its promised payments. But in addition, a loss can occur when the probability of a future credit event rises, and so the price of buying a new credit default swap for protection against that loss also rises. In this case, the firm that originally sold the credit default swap at a lower price has suffered a loss on a mark-to-market basis, and that loss is incorporated in its accounting statements. The use of mark-to-market accounting was controversial during the financial crisis (Heaton, Lucas, and McDonald 2010), but it is standard practice for most derivatives. Mark-to-market losses on AIG's credit default swap contracts were $28.6 billion in 2008 (AIG 2008, p. 265).

AIG's Credit Default Swaps

As of December 31, 2007, AIG had written credit default swaps with a notional value of $527 billion. Due to accounting conventions, the credit default swaps do not directly show up on AIG's balance sheet. These swaps were written on corporate loans ($230 billion), prime residential mortgages ($149 billion), corporate debt/collateralized loan obligations ($70 billion), and multisector collateralized debt obligations ($78 billion) (AIG 2007b, p. 122). (AIG also had an additional $1.5 trillion of other derivative exposures, including over $1 trillion in interest rate swaps.) The credit default swaps written on multisector collateralized debt obligations proved the most troublesome. Again, a collateralized debt obligation is a financial security backed by an underlying stream of debt payments, which can be from mortgages, home equity loans, credit card loans, auto loans, and other sources. The payments on this security are then divided into tranches, so that junior tranches will bear losses before senior tranches do—allowing the senior tranches to receive a higher credit rating. It is even possible to create a new collateralized debt obligation by combining tranches of other collateralized debt obligations, a so-called "CDO-squared." AIG insured collateralized debt obligations backed by a variety of assets, but including a substantial share backed by mortgages—both residential and commercial as well as prime, subprime, and Alt-A (which fall between prime and subprime on the risk spectrum) (AIG 2008, p. 139).[5] It is important to realize that AIG's credit default swap exposure resulted in a "one-way" bet on real estate: that is, a decline in real estate prices and a rise in foreclosures would impose costs on AIG, but AIG had no offsetting hedging position that would show gains if real estate prices fell. In contrast, market-making financial firms (like a stockbroker-dealer) typically seek to hedge any significant directional exposure, so that they make profits regardless of whether the price of the underlying asset (say, the price of a stock) rises or falls.

AIG (2007b, p. 122) characterized $379 billion of its credit default swaps (out of $527 billion)—those on corporate loans and prime residential mortgages—as used for "regulatory capital relief rather than risk mitigation," primarily by European banks. These do not appear to have been especially risky; in its 2008 10-K, AIG (2008, p. 118) reported a mark-to-market loss of $379 million on this portfolio, 0.1 percent of the notional value. Moreover, AIG (2007b, p. 122) expected that the swaps would be terminated by the counterparties once they were operating under the Basel II capital

[5] Details of AIG's insured multisector collateralized debt obligations and others are available online at http://fcic.law.stanford.edu/resource/staff-data-projects/cdo-Library.

rules. This suggests that the counterparty banks considered themselves compliant with Basel II, although they were not yet regulated under those rules.

AIG began originating multisector credit default swaps in 2003, at a time when the firm was rated AAA. Over half of AIG's cumulative issuances of credit default swaps, however, occurred after the firm's credit rating was downgraded twice in 2005. The AIG Financial Products subsidiary reportedly decided to stop originating credit default swaps in December 2005, at which point it still had $80 billion of commitments (Polakoff 2009, p. 5).

Collateral and Variation Margin

AIG's credit default swap contracts were traded over-the-counter—that is, directly with counterparties—as opposed to being traded on an exchange and cleared through a clearinghouse. The standard master agreement for over-the-counter derivatives is provided by the International Swaps and Derivatives Association and includes a credit support annex, which specifies how counterparty credit risk will be addressed. Both the master agreement and annex can be customized when negotiating a deal.

By construction, many derivatives contracts have zero market value at inception; this is generally true for futures, swaps, and credit default swaps. When a position has zero market value, the two parties to a contract can, by mutual consent, exit the contract without any obligation for either to make any further payment to the other. Note that one or both parties may be using the contract to hedge a position, in which case exiting would leave at least one party with some unhedged risk to consider.

As time passes and prices move, a contract initiated with zero market value will generally not remain at zero market value: fair value will be positive for one counterparty and negative by an exactly offsetting amount for the other. In such cases, it is common for the negative value party to make a compensating payment to the positive value counterparty. Such a payment is referred to as *margin* or *collateral*; in this context, the two terms mean the same thing.[6] Collateral can flow back and forth as market values change. It is important to note that this transfer of funds based on a market value change is classified as a change in collateral and not as a payment. The reason is that the contract is still active, so collateral is held by one party against the *prospect* of a loss at the future date when the contract matures or makes payment on a loss. If the contract ultimately does not generate the loss implied by the market value change, the collateral is returned. The accounting treatment of collateral recognizes this description, and the reporting of collateral on the balance sheet depends upon the existence of a master netting agreement. When full variation margin is regularly exchanged, the value of the contract is in effect regularly reset

[6] Technically, payments due to market value changes are *variation margin*. Another use of collateral is to protect against possible future market value changes. This kind of collateral, called "initial margin" or the "independent amount," was typically not used in over-the-counter markets in dealer-to-dealer transactions prior to the crisis and is not relevant for discussing AIG.

to zero, meaning that the counterparties can agree to exit the contract without any further payments.

AIG's Collateral Practices

The post-crisis investigation shed light on AIG's collateral arrangements with various counterparties. Most of the credit default swap contracts written by AIG did not call for full exchange of variation margin. Rather, they carried a wide range of collateral provisions (details are summarized in AIG 2007c, d, and market standards for collateral are discussed in ISDA 2010). Some contracts made no provision for any exchange of collateral. Most often, AIG would make collateral payments only if the decline in value of the insured assets exceeded some predefined threshold. These thresholds often depended on AIG's credit rating, which meant that a corporate ratings downgrade could lead to a large required collateral payment. Selected examples from December 2007 (AIG 2007d) illustrate agreements ranging from full mark-to-market to an 8 percent threshold with various credit rating triggers for AIG and in some cases for the underlying collateral. Here are three examples. Goldman Sachs had 44 transactions with AIG, with a total notional value of $17.09 billion. The threshold (level of market value change required to trigger a collateral payment) was "4% as long as AIGFP is rated in the AA/Aa category" (AIG 2007d, p. 4). Societe Generale had 38 transactions with AIG, with a total notional value of $18.64 billion. The threshold was "8% as long as AIGFP is rated AA/Aa2 and Reference Obligation is rated at least in the AA/Aa category; the Threshold is reduced based on a matrix that takes into account lower ratings of AIGFP and/or the Reference Obligation" (AIG 2007d, p. 6). Finally, RBS had four transactions with AIG, with a total notional value of $1.35 billion. AIG had to make variation payments for any market value change; the threshold for these was zero (AIG 2007d, p. 6).

The assets underlying the multisector collateralized debt obligations were not easily traded. As a consequence, there were running disagreements between AIG and its counterparties, later documented by the Federal Crisis Inquiry Commission, about their mark-to-market value at any given time and hence the amount of collateral that AIG owed counterparties.

Because many of the AIG credit default swap agreements did not include full payment of mark-to-market variation margin, AIG could and did accumulate unpaid losses. An unpaid variation amount is economically equivalent to a loan from the counterparty to AIG. If AIG has $1 billion in unpaid variation margin, it is as if AIG borrowed $1 billion from the counterparty. In addition, a party accumulating unpaid losses may be unwilling to exit a derivatives contract, because doing so would force it to make full collateral payments. Presumably this is why the credit support annex of swap agreements will often contain provisions that allow the purchaser of a credit default swap to terminate the agreement if the issuer of the swap experiences a credit downgrade.

AIG had first reported a loss on its written credit default swaps in 2007, losing $11.5 billion on all such swaps for the year—$11.1 billion in the fourth quarter

Table 3
Evolution of Collateral Calls and Collateral Posted for AIG's Credit Default Swaps (CDS) on Multisector Collateralized Debt Obligations (CDOs)
(millions of dollars)

	Goldman Sachs		*Societe Generale*		*Total for all counterparties*		
Date	*Call*	*Posted*	*Call*	*Posted*	*Call*	*Posted*	*Total shortfall*
6/30/2008	7,493	5,913	1,937	1,937	15,780	13,241	2,539
9/12/2008	8,979	7,596	4,280	4,008	23,441	18,922	4,519
9/15/2008[a]	10,072	7,596	9,833	4,320	32,013	19,573	12,440
9/16/2008	10,065	7,596	9,818	5,582	33,879	22,445	11,434

Source: "AIG/Goldman Sachs Collateral Call Timeline," Financial Crisis Inquiry Commission (FCIC). http://fcic.law.stanford.edu/documents/view/2172.
[a] AIG was downgraded on September 15, 2008, and this meant that many multisector CDS counterparties were contractually entitled to additional collateral.

alone—with 98 percent of the total coming from credit default swaps on multisector collateralized debt obligations (AIG 2007b, p. 83).[7] Losses continued in 2008. Table 3 depicts the evolution of collateral calls between June and September 2008 for Goldman Sachs and Societe Generale (AIG's two largest credit default swap counterparties), as well as for all counterparties combined. As of June 30, 2008, counterparties had called $15.78 billion and AIG had posted $13.24 billion. The totals climbed gradually until on September 12, 2008, total calls amounted to $23.44 billion, with AIG having posted $18.92 billion. Thus, prior to the rescue, AIG had already provided almost $20 billion to counterparties.

The effect of triggers from changes in credit ratings is evident in a comparison of collateral calls for September 12, 2008, and those for September 15, 2008, the day on which all three credit ratings agencies downgraded AIG below AA−. Total collateral calls increased by $8.6 billion, to $32 billion. AIG's collateral shortfall rose from $4.5 billion to $12.4 billion. Societe Generale's call on that day rose by $5.5 billion.

What Would Have Happened to Credit Default Swap Counterparties If AIG Had Declared Bankruptcy?

If AIG had declared bankruptcy on September 16, 2008, what would have been the direct effect on credit default swap counterparties? It is of course impossible to answer this question definitively, but some straightforward observations are possible.

[7] AIG's credit default swap business was barely disclosed prior to 2007. The phrase "super senior" referring to tranches of collateralized debt obligations appears four times in the 2006 annual report and 114 times in 2007; "multisector" does not appear in 2006, but appears 23 times in 2007; "CDO" (for

Table 4

Multisector Credit Defalt Swap (CDS) Counterparty Collateral Shortfall Relative to Equity and Asset Sales Necessary to Maintain Pre-shortfall Equity-to-Asset Ratio

	Total assets *($ billions)* [1]	*Total shareholders equity* *($ billions)* [2]	*AIG shortfall as of 9/16/2008* *($ billions)* [3]	*Shortfall/ equity* *[3]/[2]* [4]	*Asset sales to return to pre-AIG-shortfall equity-to-assets ratio* *($ billions)* [5]
Goldman Sachs	1,081.8	45.6	2.5	5.41%	58.5
Societe Generale	1,694.4	56.0	4.2	7.56%	128.1
Merrill Lynch	875.8	38.4	1.0	2.70%	23.6
UBS	1,784.5	41.5	1.0	2.41%	43.0
DZ Bank	677.0	10.6	0.7	7.00%	47.4
Rabobank	894.0	45.0	0.6	1.31%	11.7
Total					**312.4**

Source: Federal Crisis Inquiry Commission "AIG/Goldman-Sachs Collateral Call Timeline," available at http://fcic.law.stanford.edu/documents/view/2172 and author calculations using 2008 Q2 and Q3 financials. Goldman Sachs, Merrill Lynch, and UBS assets, shareholders equity, and tier 1 capital come from 2008Q3 financial statements. Societe Generale, DZ Bank, and Rabobank values come from 2008Q2 financial statements. For each counterparty, to get the number shown in column 5, multiply total assets shown in column 1 by the percentage shown in column 4. Column 5 represents the assets sales that would be necessary if the AIG collateral shortfall from column 3 was realized and the firm in question chose to preserve its original equity-to-asset ratio.

AIG had 21 counterparties for its multisector credit default swaps. Of those, nine had collateral calls exceeding $500 million, and six of those—Goldman Sachs, Societe Generale, Merrill, UBS, DZ Bank, and Rabobank—had a difference greater than $500 million between the collateral they had requested and the amount AIG had posted. Table 4 shows these collateral shortfalls for the six largest counterparties to AIG's multisector credit default swaps as of September 16, 2008, and also shows the shortfall relative to shareholder equity for each counterparty. Of the $11.4 billion that AIG owed to counterparties on its credit default swaps on September 16, 2008, these six banks accounted for $10 billion.

If AIG had defaulted, the counterparty banks to the credit default swaps on the multisector collateralized debt obligation would have likely faced three direct consequences. First, the banks would have kept the collateral already posted by AIG. This is a result of the rule mentioned earlier that derivatives are exempted from the automatic stay in bankruptcy (for discussion, see Edwards and Morrison 2005;

collateralized debt obligation) appears twice in 2006 and 93 times in 2007. AIG's 2006 annual report discloses that it had written $483.6 billion in credit default swaps, but provides no details, whereas the 2007 report reports notional values of credit default swap by category. AIG's first public disclosure of credit default swaps written on the multisector collateralized debt obligations came on August 9, 2007, during a second-quarter earnings call (Federal Crisis Inquiry Commission 2011, p. 268). The lack of disclosure is surprising given that the credit default transactions increased the size of AIG's balance sheet by 50 percent in economic terms.

Bolton and Oehmke forthcoming). Second, the banks would have been treated as general creditors for any collateral that had been requested but AIG had not yet posted. Third, the banks would have retained the asset or position that had been hedged by the defaulted credit default swap.

Assuming that assets were valued correctly and that the September 15, 2008, downgrade of AIG to an A rating eliminated any remaining thresholds that might have further increased collateral calls, the economic cost of an AIG default for its counterparties would be equal to the collateral shortfall: that is, the difference between called and posted collateral. How significant would this shortfall have been for the counterparty banks? As can be seen in Table 4, even for the six banks that were individually owed more than $500 million, in no case did the shortfall exceed 10 percent of their equity capital.

However, comparing the actual loss with counterparty equity may be too sanguine, because it assumes that counterparties would simply absorb the loss. This assumption faces at least three potential problems. First, Brunnermeier and Pedersen (2009) and Duarte and Eisenbach (2014), among others, emphasize the possibility of fire-sale spillovers. Institutions might respond to the loss in capital by selling assets in order to return to their pre-loss leverage ratios. This could lower asset prices and lead to mark-to-market losses at other firms who might in turn sell assets to get back to target leverage ratios. Our back-of-the-envelope calculations presented in Table 4 suggest that if these six banks had chosen to respond by selling assets to get back to their pre-AIG default debt to equity ratios, they would have needed to sell $312 billion in assets. Second, the cancellation of the credit default swaps would leave many of the counterparties with unhedged exposure to real estate risk. Retaining this risk could reduce the capacity for other risk-taking. Third, even if one concludes that counterparties could have absorbed losses due to an AIG failure, other market participants would not have known at the time who was exposed and in what amount. For this reason, the failure of any large financial firm can be stressful for the financial system—a conclusion that is not particular to credit default swaps or AIG.

Another consequence of AIG's failure would have been cancellation of the $387 billion of other credit default swaps mainly held by European banks. Collateral calls related to these positions totaled just $500 million on September 16, 2008 (Congressional Oversight Panel 2010, p. 42), and as noted above, the institutions were apparently anticipating the swaps to expire when they adopted Basel II capital rules. The cancellation of these swaps would have created a regulatory capital deficiency, but it is not clear that this would have been economically important. In any event, European financial regulators would have had the option to forebear from enforcing the capital rules for a time, thus allowing for a period of adjustment.

Overall, how much did the rescue of AIG benefit its multisector credit default counterparties? Some media reports suggest that $62 billion in taxpayer funds were paid to AIG's multisector credit default swap counterparties (for example, Orol 2010). In fact, the direct counterparty benefit from the rescue is smaller. We can divide the payments to AIG's credit default swap counterparties into three categories.

First, there are collateral payments AIG made prior to the rescue. These payments would have been retained by counterparties in a bankruptcy and therefore cannot be attributed to the rescue. These payments totaled $22.4 billion with $18.5 billion associated with multisector collateralized debt obligations that became part of the Maiden Lane III Fed-created special purpose vehicle (see also Congressional Oversight Panel 2010, p. 93). Second, there are collateral payments made by AIG after the rescue. These payments could only be made because of the rescue and clearly offset losses that counterparties would have sustained in the absence of a rescue. This amount provides a lower bound on the assistance received by counterparties to the credit default swaps due to the rescue. AIG's 2008 10-K reports total collateral payments for credit default swaps of $40.1 billion for 2007 and 2008, suggesting that $17.7 billion was paid after the rescue. (As confirmation of this amount, the Congressional Oversight Panel (2010, p. 93) found that collateral payments of $16.5 billion were made after the rescue for the assets that became part of Maiden Lane III.) Finally, Maiden Lane III made cash payments of $26.8 billion in exchange for the assets that AIG had insured. These payments were equal to the estimated fair market value of the assets at the time (Office of the Special Inspector General 2009). While there may not have been many buyers for these assets, even at 47 percent of face value in the fall of 2008, it is inappropriate to consider the entire amount of the price that Maiden Lane III paid for the credit default swap as a direct benefit to the counterparties. Indeed, as we discuss in the next section, this portfolio of assets appreciated and was later sold for a modest gain.

Performance of Maiden Lane Assets

The Federal Reserve Bank of New York created several special purpose vehicles as part of the rescue of AIG. Among them, Maiden Lane II purchased the remaining securities lending invested collateral from AIG, and Maiden Lane III acquired from AIGFP's counterparties the collateralized debt obligations that AIG had insured. This acquisition terminated the associated credit default swaps. Maiden Lane II was funded by a $19.5 billion loan from the New York Fed and $1 billion from AIG that would absorb the first $1 billion in losses. Maiden Lane III was funded by a loan from the New York Fed of $24.3 billion and $5 billion in equity from AIG (Congressional Oversight Panel 2010, pp. 87, 91). The New York Fed has thoroughly documented the resulting cash flows at http://www.newyorkfed.org/markets/maidenlane.html. These data, in combination with information from various other sources, allow us to examine how the value of these securities evolved both while they were held in the Maiden Lane vehicles and afterward.

Maiden Lane II and III Performance

The New York Fed managed the Maiden Lane vehicles and assets with the goal of selling the assets once markets stabilized. Both Maiden Lane vehicles were ultimately liquidated for a total gain of $9.5 billion. While held in the Maiden Lane vehicles, the underlying securities paid interest and also repaid principal and

Table 5

Summary Statistics for Assets in Maiden Lane II and Maiden Lane III Portfolios

	Maiden Lane II assets			*Maiden Lane III assets*		
	Min.	*Median*	*Max.*	*Min.*	*Median*	*Max.*
Notional (millions $)	0.02	31.00	266.00	0.04	201.00	5,400.00
Purchase percentage	0.01	0.56	0.99	0.10	0.48	0.94
Sale percentage	0.00	0.58	1.02	0.03	0.49	0.96
Gain (millions $)	−70.50	1.53	76.40	−172.00	36.80	779.00
Return (Gain/Purchase Price − 1)	−0.95	0.13	4.06	−0.85	0.35	1.24
Benchmark return	−0.15	0.22	0.23	0.03	0.21	0.23
Return less Benchmark return	−1.18	−0.07	3.84	−0.91	0.14	1.02

Source: Authors' calculations using data from the Federal Reserve Bank of New York and Markit.
Notes: "Purchase percentage" is the ratio of the price paid for each asset to its notional value. "Sale percentage" is the ratio of the price received for each asset to its notional value. The "Benchmark return" for Maiden Lane II is the return on the ABX.HE.AAA.06-1, an index of AAA-securitized subprime mortgage loans originated in the last six months of 2005. For Maiden Lane III the "benchmark return" is 70 percent ABX.HE.AAA.06-1 and 30 percent CMBX.NA.AAA.1-1, an index of commercial mortgage-backed obligations.

experienced write-downs, both of which reduced their face value. They were ultimately sold by auction. The Maiden Lane II assets were bought in December 2008 for $20.5 billion (53 percent of par value), returned $8.9 billion in interest and principal while held, and the residual claims were sold for $15.1 billion (51 percent of par) for a nonannualized return of 16.9 percent. The securities were sold principally in 2011 and 2012. Table 5 summarizes the size, purchase and sale discount, and returns of the individual Maiden Lane II and III securities. There is significant variation in the size and discounts of securities.

It is not obvious whether the overall return of 16.9 percent is "good," given the risk of the assets. We can ask, however, whether the Maiden Lane securities performed especially well or poorly compared to a broader universe of residential real estate. To perform this comparison while controlling for different liquidation dates, we use as a benchmark an index of AAA-securitized subprime mortgage loans originated in the last six months of 2005, the ABX.HE.AAA.06-1 index. The median security in Maiden Lane II had a 13 percent return and underperformed the ABX by 7 percent. It is worth noting that AIG had begun to sell its securities lending collateral prior to the creation of Maiden Lane II, and the securities acquired by the special purpose vehicle were likely the poorest assets.

The securities in Maiden Lane III—primarily the multisector collateralized debt obligations that AIG had insured through its credit default swaps—were bought in November and December 2008 for $29.3 billion (47 percent of par), returned $17.1 billion in interest and principal, and were sold for $22.6 billion (50 percent of par), for a nonannualized return of 35.1 percent. The securities were sold primarily in 2012. The median security in Maiden Lane III returned 35 percent, exceeding the

Table 6

Aggregate Performance of Maiden Lane Asset: Origination through October 31, 2014

	Date			
	At origination	*Beginning of Maiden Lane*	*Maiden Lane sale*	*Most recent*
ML2 notional (billions)	$137.7	$85.9	$62.6	$43.2
ML2 amortization (billions)	$0.00	$51.8	$72.6	$87.4
ML2 write-down (billions)	$0.00	$0.05	$2.5	$7.0
ML2 write-down since start (%)	**0.00%**	**0.04%**	**1.8%**	**5.1%**
ML2 securities with write-downs (%)	**0.00%**	**0.5%**	**17.5%**	**36.0%**
ML3 notional (billions)	$82.5	$68.8	$45.8	$29.5
ML3 amortization (billions)	$0.00	$13.7	$31.0	$43.1
ML3 write-down (billions)	$0.00	$0.00	$5.7	$9.9
ML3 write-down since start (%)	**0.00%**	**0.00%**	**6.9%**	**12.0%**
ML3 securities with write-downs (%)	**0.00%**	**0.00%**	**47.2%**	**59.0%**

Source: Authors' calculations based on data from the Federal Reserve Bank of New York and from summaries derived from Intex data. Analysis using the Intex data was performed by Larry Cordell and Yilin Huang of the Federal Reserve Bank of Philadelphia.

Notes: Data were available for each of the 855 securities in Maiden Lane II and 146 of the 155 securities in Maiden Lane III, accounting for 97 percent of the original Maiden Lane III face amount. Omitted securities were either not present in the Intex data (seven securities) or had partially missing data (two securities). "Origination" is the date the security was created; "Beginning of Maiden Lane" is the approximate time at which the asset was purchased by a Maiden Lane; "Maiden Lane Sale" is the approximate time at which the asset was a sold by a Maiden Lane; and "Most Recent" refers to information as of October 31, 2014 or the most recent prior data available. (Some assets matured or were written down completely prior to October 31, 2014. Once a security has been paid off or written down completely, no additional data are reported for it.) Figures reflect the full outstanding amount for any security that was included in Maiden Lane II or III and not the share of the security purchased by those vehicles. For example, Maiden Lane II might have owned 10 percent of a particular security and 100 percent of the outstanding amount of the security is used to compute the figures in the table.

benchmark return by 14 percent. Returns on the Maiden Lane III securities were greater than those on Maiden Lane II, even after adjusting for the return benchmark. (The benchmark for Maiden Lane III was 70 percent ABX.HE.AAA.06-1 and 30 percent CMBX.NA.AAA.1-1, an index of commercial mortgage backed obligations. We obtained almost identical results using this benchmark and using ABX alone.)

Post–Maiden Lane Performance

Table 6 shows the performance of the securities lending invested collateral portfolio that eventually became part of Maiden Lane II and the super senior tranches of the collateralized debt obligations that were insured by AIGFP and eventually became part of Maiden Lane III.[8] The table provides information at

[8] Figures reported in Table 6 reflect the full outstanding amount for any security that was included in Maiden Lane II or III and not the share of the security purchased by those vehicles. Please see the notes to Table 6 for additional details.

four points: when the securities were originated (various dates); when the Maiden Lane vehicles were created; when the securities were sold from the Maiden Lane vehicles (various dates); and as of October 2014 (or the most recent prior date for which information is available). Thirty-six percent of the Maiden Lane II securities and 59 percent of the Maiden Lane III securities in the table have experienced write-downs. A sizeable share of write-downs have occurred during the post–Maiden Lane period. As explained earlier, senior tranches will be the last to experience actual losses, and for this reason, actual losses in these tranches will appear later and will likely increase over time. With approximately one-third of principal still outstanding, future substantial writedowns for the assets in both Maiden Lanes II and III remain possible.

Reported write-downs to date are 5.1 percent of the original face value of the securities that ended up in Maiden Lane II and 12 percent for Maiden Lane III. These estimates were calculated from information provided by Larry Cordell and Yilin Huang from the Federal Reserve Bank of Philadelphia, following the methodology in Cordell, Huang, and Williams (2011). The Maiden Lane III assets are harder to assess because issuers of collateralized debt obligations do not report writedowns prior to maturity. It is thus necessary to look for writedowns on the individual instruments constituting the collateralized debt obligation. The fact that the Maiden Lane II and III assets have suffered write-downs means that we can reject the stark claim that they were "money good."

Was AIG Special?

Given the drama surrounding AIG, it is natural to ask how AIG compared to other financial firms at the time. Was AIG unusual in its risk-taking or was it just unlucky? It turns out that AIG resembled some large banks in important respects: its real estate holdings were comparable to those of Citigroup and Bank of America, banks which also received considerable official support in 2008 and 2009. In addition, AIG's financing of its real estate positions was fragile and prone to runs in times of financial difficulty. Making a comparison with other firms requires first that we assess AIG's position prior to the rescue, especially its exposure to housing. A notable feature of AIG was its large position in written credit default swaps and we need to take these into account when comparing firms.

Issuing a credit default swap is economically equivalent to borrowing in order to finance the purchase of the same risky bond that the credit default swap would insure. To see this, suppose that you have excellent credit, that you borrow \$50 at a 5 percent rate of interest, and that you use the proceeds to buy \$50 in one-year bonds that might default, and which consequently pay a 15 percent rate of interest. If the bonds pay in full, you have a \$57.50 asset ($50 + .15 \times 50 = 57.50$), offset by a \$52.50 liability ($50 + .05 \times 50 = 52.50$), and you will have earned the 10 percent interest differential (\$5). However, if the bonds lose \$20, for example, you have a \$30 asset and a \$52.50 liability—and you face a loss of \$22.50. This pattern of gains and losses is precisely that faced by the seller of a credit default swap on the bonds. If the bonds pay in full, the seller earns the credit default swap premium ($5), and if the bonds default, the credit default swap seller bears the loss ($22.50) that is paid to the bondholder.[9]

To relate this insight to AIG, consider the simplified example of a firm with $100 in assets—$90 of debt and therefore $10 of equity. The firm has an asset-to-equity ratio of 10:1 (that is, $100/$10). This firm now sells a credit default swap on $50 of mortgage-backed securities. In the contract, the buyer of the credit default swap agrees to make an annual payment of $5, and the seller bears the loss if the mortgage-backed securities fail. The economic result is the same as if the firm had $150 in assets ($100 plus the $50 in mortgage-backed securities insured by the credit default swap), financed with $140 in debt, $50 of which is implicit in the credit default swap. The issuance of a credit default swap implicitly changes assets and debt, but not equity.

This was approximately AIG's situation: the firm as a whole had $1.06 trillion of assets and about $964 billion in liabilities at the end of 2007, so it had equity of $96 billion. It issued $527 billion in credit default swaps. It was therefore economically equivalent to a firm with $1.59 trillion in assets and $96 billion in equity. Taking into account the credit default swaps, AIG's ratio of assets to equity was 16:1 rather than 11:1.

AIG was not the only financial firm with off-balance sheet real estate holdings. Citigroup, Bank of America, and JPMorgan Chase all had off-balance-sheet asset-backed commercial paper conduits used to fund real estate holdings (Acharya, Schnabl, and Suarez 2013). The effective asset-to-equity ratio for these banks was also higher than reported.

Using these insights, we compared AIG's total real estate exposure with Citigroup, Bank of America, and JPMorgan Chase and with that of another large insurance company, Metlife. Our calculations appear in an online Appendix available with this paper at http://e-jep.org, in Appendix Table X1. After adjusting the balance sheets as discussed above, we find that AIG's real estate exposure was 24 percent of assets, comparable to that of Bank of America (32 percent) and Citigroup (21 percent). AIG's effective real estate holdings were almost four times its book equity.

Was AIG effectively acting like a bank? Banks typically employ short-term financing to fund holdings of long-term illiquid assets. AIG did have some explicit short-term financing, in particular $20 billion of commercial paper. But AIG's illiquid real estate positions were also financed in a way that was not as transparently fragile as demand deposits, but which could create large liquidity needs if AIG suffered losses.

[9] In economic terms, a credit default swap is economically equivalent to a purchase of the insured asset financed by issuing floating rate debt (Duffie 1999). For a general discussion of credit default swaps, see McDonald (2013, chap. 27).

As discussed earlier, AIG's securities lending agreements had a relatively short maturity and could be subject to early termination. As AIG suffered downgrades and as the real estate investments made with securities lending proceeds suffered losses, securities lending counterparties became increasingly likely to terminate these agreements, culminating in a $5.2 billion redemption request on September 15, 2008. This desire by counterparties to unwind their exposure to AIG resembled a bank run, as counterparties sought to unwind the positions rather than be left with collateral and possibly involved in lawsuits. AIG effectively used collateralized short-term financing to buy real estate assets.

Although the mechanism was different, AIG's multisector credit default swap positions also suffered from something akin to a bank run. AIG's credit default swap counterparties could not unilaterally terminate credit default swap agreements, but they were entitled to collect collateral as the values of insured assets declined and these counterparty rights could sometimes be accelerated if AIG's credit rating was lowered. When AIG was downgraded on September 15, 2008, collateral calls on AIG's multisector credit default swaps increased by $8.6 billion as a result. Thus, while AIG was not literally a bank, it undeniably had bank-like characteristics as it employed financing (both explicit and implicit) that was subject to termination and cash demands when asset values fell.

Conclusions

Insurance companies are traditionally less vulnerable to financial crises than banks, in large part because they have relatively low-risk assets and do not rely heavily on short-term funding. However, AIG made itself vulnerable in a number of ways. Notably, AIG's near-failure was a result of two outsized bets on real estate, both of which generated large needs for liquidity. First, AIG used securities lending to transform insurance company assets into residential mortgage-backed securities and collateralized debt obligations, ultimately losing $21 billion and threatening the solvency of its life insurance subsidiaries. On one day in 2008, AIG was required to pay $5.2 billion in cash to satisfy redemption requests. Second, AIG issued credit default swaps on real estate–backed multisector collateralized debt obligations, ultimately losing more than $30 billion and facing a one-day $8.6 billion collateral demand due to a downgrade in its credit rating. Securities lending and writing credit default swaps were both "carry trades:" that is, bets that long-term assets would earn a higher return than the short-term cost of funding. AIG's use of financial markets to transform itself from a traditional insurance company to a bank-like firm ultimately proved disastrous.

The rescue of AIG had many beneficiaries. The broader financial system was spared the unpredictable consequences of a large and complicated firm failing at a time when financial markets were very fragile. Direct beneficiaries of the rescue included the life insurance subsidiaries that received $20 billion in capital infusions, protecting their policyholders. The counterparties to the credit fault swaps

AIG had sold on multisector credit default obligations (CDOs) were also beneficiaries, although their direct benefit was the $17.7 billion in collateral payments made after the rescue rather than much larger figures that sometimes have been emphasized. In addition to addressing problems with securities lending and the multisector credit default swap portfolio, rescue funds provided to AIG directly benefited numerous other counterparties including AIG's employees, holders of AIG's commercial paper and other AIG debt holders and repo counterparties, states and municipalities who had AIG-sponsored Guaranteed Investment Agreements, as well as defined contribution pension plans holding stable "value wraps" (which smooth the volatility of the pension plan) issued by AIG.

AIG's near failure is often described as a liquidity event: that is, it found itself in 2008 holding a number of mortgage-based securities that were impossible to sell—except perhaps at unreasonably low "fire sale" prices. But AIG sustained a loss of $99 billion in 2008, exceeding the firm's end-of-2007 equity of $96 billion (AIG 2008, p. 36), raising the question of whether it experienced a liquidity problem, a solvency problem, or both. Despite its reliance on fragile sources of funding, AIG had no specialized liquidity risk committee until 2007 (AIG 2007b, p. 99). It is tempting to attribute this to the company's insurance origins together with the belief of senior management that the real estate-related investments were "money good." Our examination of the performance of AIG's underlying real estate securities indicates that AIG's problems were not purely about liquidity. While we cannot say whether prices in 2008 were "correct" in any meaningful sense, the assets represented in both Maiden Lane vehicles have experienced substantial write-downs, with the possibility of more in the future. With hindsight, it may seem obvious that AIG's real estate assets were not "money good" and would suffer real losses, but the belief that they would not, and that liquidity would not be a problem, was an important factor in their creation and purchase by AIG and others.

■ *David Autor, Ben Chabot, Larry Cordell, Mark Finn, Scott Frame, Chiang-Tai Hsieh, Yilin Huang, Arvind Krishnamurthy, Anil Kashyap, Andreas Lehnert, Debbie Lucas, David Marshall, Richard Miller, Richard Rosen, David Scharfstein, Robert Steigerwald, and Timothy Taylor provided helpful commentary and feedback, as did seminar participants at Case Western and the Federal Reserve Banks of New York and Chicago. We are grateful to Kyal Berends, Mike Mei, and Thanases Plestis for excellent research assistance. The views presented here are solely our own and do not reflect those of the Federal Reserve Bank of Chicago or the Board of Governors of the Federal Reserve System.*

References

Acharya, Viral V., **Douglas Gale**, **and Tanju Yorulmazer.** 2011. "Rollover Risk and Market Freezes." *Journal of Finance* 66(4): 1177–1209.

Acharya, Viral, Philip Schnabl, and Gustavo Suarez. 2013. "Securitization without Risk Transfer." *Journal of Financial Economics* 107(3): 515–36.

Adrian, Tobias, Brian Begalle, Adam Copeland, and Antoine Martin. 2013. "Repo and Securities Lending." Federal Reserve Bank of New York Staff Report 529.

Aggarwal, Reena, Pedro A. C. Saffi, and Jason Sturgess. 2012. "Role of Institutional Investors in Voting: Evidence from the Securities Lending Market." Georgetown McDonough School of Business Research Paper No. 2012-07.

American International Group (AIG). 2007a. "Form 10-Q Quarterly Report for the Quarterly Period Ended September 30, 2007." *AIG Investor Relations.* http://services.corporate-ir.net/SEC/Document.Service?id=P3VybD1hSFIwY0RvdkwyRndhUzUwWlc1cmQybDZZWEprTG1OdmJTOWtiM2R1Ykc5aFpDNXdhSEEvWVdOMGFXOXVQVkJFUmlacGNHRm5aVDAxTWpZd05EQTRKbk4xWW5OcFpEMDFOdz09JnR5cGU9MiZmbj1BbWVyaWNhbkludGVybmF0aW9uYWxHcm91cF8xMFFfMjAwNzExMDcucGRm.

American International Group (AIG). 2007b. "Form 10-K Annual Report for the Fiscal Year ended December 31, 2007." *AIG Investor Relations.* http://services.corporate-ir.net/SEC/Document.Service?id=P3VybD1hSFIwY0RvdkwyRndhUzUwWlc1cmQybDZZWEprTG1OdmJTOWtiM2R1Ykc5aFpDNXdhSEEvWVdOMGFXOXVQVkJFUmlacGNHRm5aVDAxTlRBd016azVKbk4xWW5OcFpEMDFOdz09JnR5cGU9MiZmbj1BbWVyaWNhbkludGVybmF0aW9uYWxHcm91cF8xMEtfMjAwODAyMjgucGRm.

American International Group (AIG). 2007c. "AIG Status of Collateral Call Postings." *FCIC Resource Library.* http://fcic-static.law.stanford.edu/cdn_media/fcic-docs/2007-12-31_AIG_Status_of_Collateral_Call_Postings.pdf.

American International Group (AIG). 2007d. "AIG Super Senior Credit Transactions—Principal Collateral Provisions." *FCIC Resource Library.* http://fcic-static.law.stanford.edu/cdn_media/fcic-docs/2007-12-07%20AIG%20Super%20Senior%20Credit%20Transactions-%20Principal%20Collateral%20Provisions.pdf.

American International Group (AIG). 2008. "Form 10-K Annual Report for the Fiscal Year ended December 31, 2008." *AIG Investor Relations.* http://services.corporate-ir.net/SEC/Document.Service?id=P3VybD1hSFIwY0RvdkwyRndhUzUwWlc1cmQybDZZWEprTG1OdmJTOWtiM2R1Ykc5aFpDNXdhSEEvWVdOMGFXOXVQVkJFUmlacGNHRm5aVDAyTVRjek1qZ3dKbk4xWW5OcFpEMDFOdz09JnR5cGU9MiZmbj1BbWVyaWNhbkludGVybmF0aW9uYWxHcm91cF8xMEtfMjAwOTAzMDIucGRm.

American International Group Investor Meeting. Final. 2007. December 5. http://www.fcic.gov/documents/view/1139.

Bank of England. 2010. "Securities Lending: An Introductory Guide." http://www.bankofengland.co.uk/markets/gilts/sl_intro_green_9_10.pdf.

Bernanke, Ben S. 2009. Testimony Before the Committee on Financial Services U.S. House of Representatives, Washington, DC, March 24, 2009.

Bolton, Patrick, and Martin Oehmke. Forthcoming. "Should Derivatives Be Privileged in Bankruptcy?" *Journal of Finance.*

Brunnermeier, Markus. K., and Lasse Heje Pedersen. 2009. "Market Liquidity and Funding Liquidity." *Review of Financial Studies* 22(6): 2201–38.

Congressional Oversight Panel. 2010. "June Oversight Report: The AIG Rescue, Its Impact on Markets, and the Government's Exit Strategy." June 10. http://cybercemetery.unt.edu/archive/cop/20110402010341/http://cop.senate.gov/documents/cop-061010-report.pdf.

Cordell, Larry, Yilin Huang, and Meredith Williams. 2011. "Collateral Damage: Sizing and Assessing the Subprime CDO Crisis." Federal Reserve Bank of Philadelphia Working Paper 11-30.

Dinallo, Eric. 2010. "What I Learned at the AIG Meltdown: State Regulation Wasn't the Problem." *Wall Street Journal*, February 2. http://online.wsj.com/articles/SB10001424052748704022804575041283535717548.

Duarte, Fernando, and Thomas M. Eisenbach. 2014. "Fire-Sale Spillovers and Systemic Risk." Federal Reserve Bank of New York Staff Report 645.

Duffie, Darrell. 1999. "Credit Swap Valuation." *Financial Analysts Journal* 55(1): 73–87.

Edwards, Franklin R., and Edward R. Morrison. 2005. "Derivatives and the Bankruptcy Code: Why the Special Treatment?" *Yale Journal on Regulation* 22: 91–122.

Ellul, Andrew, Chotibhak Jotikasthira, Christian T. Lundblad, and Yihui Wang. 2014. "Mark-to-Market Accounting and Systemic Risk: Evidence from the Insurance Industry." *Economic Policy* 29(78): 297–341.

Federal Deposit Insurance Corporation (FDIC).

2011. "The Orderly Liquidation of Lehman Brothers Holdings Under the Dodd–Frank Act." *FDIC Quarterly* vol. 5, no. 2. https://www.fdic.gov/bank/analytical/quarterly/2011_vol5_2/lehman.pdf.

Financial Crisis Inquiry Commission (FCIC). 2011. *The Financial Crisis Inquiry Report: Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States. FCIC Resource Library*. http://fcic-static.law.stanford.edu/cdn_media/fcic-reports/fcic_final_report_full.pdf.

Fitch Ratings. 2006. "Legal Status of Derivative Counterparties in U.S. Insurance Company Insolvencies Remains Murky, Increases Risk."

Government Accountability Office. 2011. *Review of Federal Reserve System Financial Assistance to American International Group, Inc.* http://www.gao.gov/products/GAO-11-616.

Heaton, John C., Deborah Lucas, and Robert L. McDonald. 2010. "Is Mark-to-Market Accounting Destabilizing? Analysis and Implications for Policy." *Journal of Monetary Economics* 57(1): 64–75.

Hutchings, Mark. 2010. "FCIC Staff Audiotape of Interview with Mark Hutchings, American International Group, Inc." Financial Crisis Inquiry Commission, June 22. http://fcic-static.law.stanford.edu/cdn_media/fcic-audio/2010-06-22%20FCIC%20staff%20audiotape%20of%20interview%20with%20Mark%20Hutchings,%20American%20International%20Group,%20Inc.mp3.

International Swaps and Derivatives Association (ISDA). 2010. "Market Review of OTC Derivative Bilateral Collateralization Practices." March 1. http://www.isda.org/c_and_a/pdf/Collateral-Market-Review.pdf.

Kacperczyk, Marcin, and Philipp Schnabl. 2010 "When Safe Proved Risky: Commercial Paper during the Financial Crisis of 2007–2009." *Journal of Economic Perspectives* 24(1): 29–50.

Krishnamurthy, Arvind, Stefan Nagel, and Dmitry Orlov. 2014. "Sizing Up Repo." *Journal of Finance* 69(6): 2381–2417.

McDonald, Robert L. 2013. *Derivatives Markets,* 3rd ed. Boston, MA: Pearson/Addison Wesley.

Office of the Special Inspector General for the Troubled Asset Relief Program. 2009. "Factors Affecting Efforts to Limit Payments to AIG Counterparties." November 17. SIGTARP-10-003.

Orol, Ronald D. 2010. "Geithner, Paulson Defend $182 Billion AIG Bailout." *MarketWatch*, January 27. http://www.marketwatch.com/story/geithner-paulson-defend-182-bln-aig-bailout-2010-01-27.

Pierce, Hester. 2014. "Securities Lending and the Untold Story in the Collapse of AIG." George Mason University, Mercatus Center Working Paper No. 14-12.

Polakoff, Scott M. 2009. Statement of Scott M. Polakoff, Acting Director, Office of Thrift Supervision, regarding "American International Group: Examining What Went Wrong, Government Intervention, and Implications for Future Regulation" before the Committee on Banking, Housing, and Urban Affairs, US Senate, March 5, 2009. Available at: http://www.occ.gov/news-issuances/ots-testimony-archive.html.

Risk Managing Association. 2007. "Securities Lending: Quarterly Aggregate Composite."

Taibbi, Matt. 2011. *Griftopia*. Random House.

Legal, Political, and Institutional Constraints on the Financial Crisis Policy Response[†]

Phillip Swagel

As the financial crisis manifested itself and peaked in 2007 and 2008, the response of US policymakers and regulators was shaped in important ways by legal and political constraints. Policymakers lacked certain legal authorities that would have been useful for addressing the crisis, notably to use public capital to stabilize the banking sector or to deal with the failure of large financial firms such as insurance companies and investment banks that were outside the scope of bank regulators' authority to resolve deposit-taking commercial banks. US policymakers had long been aware that new legal authorities might be useful and even necessary, but political constraints meant that such changes could only be enacted after a financial market crisis actually threatened the economy. Analyzing the response to the crisis and considering improvements to future efforts thus requires understanding the political and legal constraints that narrowed the available options or affected the timing of actions taken.

Legal constraints were keenly felt at the US Department of the Treasury, where I served as a senior official from December 2006 to January 2009. Treasury had virtually no emergency economic authority at the onset of the crisis in 2007, with the exception of the Treasury's Exchange Stabilization Fund, which was intended for use in exchange rate interventions. Even while options such as the capital injections ultimately undertaken through the Troubled Asset Relief Program (the TARP) were being developed at the Treasury in spring 2008, policymakers felt that it was possible

■ *Phillip Swagel is Professor in International Economic Policy, University of Maryland School of Public Policy, College Park, Maryland. From December 2006 to January 2009, he was Assistant Secretary for Economic Policy at the US Department of the Treasury. His email address is pswagel@umd.edu.*

[†] To access the disclosure statement, visit http://dx.doi.org/10.1257/jep.29.2.107

doi=10.1257/jep.29.2.107

to propose the necessary changes in the law to authorize the response only when the Secretary of the Treasury and the Chairman of the Federal Reserve could tell Congress that action was necessary to avoid an economic collapse. This constraint explains why, as the systemic risks of the financial crisis became apparent, the initial policy response largely fell to the Federal Reserve, which had the authority to act under emergency circumstances.

The story of the financial crisis response can be told through the lens of evolving legal and political constraints. In late 2007 and early 2008, while policymakers recognized weaknesses in the system, they believed that conventional monetary and fiscal responses such as Fed lending and a modest fiscal stimulus would suffice to buoy the US economy while the imbalances that had built up during the housing bubble were resolved (indeed, Broda and Parker 2014 show that the early 2008 stimulus increased consumption). By the time of the Bear Stearns bailout in March 2008, the usual methods were clearly perceived to be inadequate, and the Fed was making discretionary choices to invoke authority reserved for "unusual and exigent" circumstances to respond to the potential collapse of a nonbank financial firm. In September 2008, the Fed's ability to use this discretionary authority had reached its limits, and the imminent risk of financial crisis led to the Troubled Asset Relief Program, which authorized public money to be used to purchase troubled assets such as subprime mortgage-backed securities from banks or to inject capital into the banking system by purchasing shares of preferred stock in banks. The advent of the TARP capital injections facilitated a program of guarantees by the Federal Deposit Insurance Corporation to support bank funding, undertaken with existing legal authority but in an extraordinary way. Together, these actions reassured market participants that the US financial sector would not collapse and marked the beginning of the stabilization from the crisis.

There will inevitably be another financial crisis, and the response will be shaped by both the lessons learned from recent history and the statutory and political changes in the wake of the crisis. The paper thus concludes by discussing changes in constraints since the crisis, with a focus on two developments: 1) the political reality that there will not in the near future be another wide-ranging grant of fiscal authority as was given with the Troubled Asset Relief Program, and 2) the new legal authorities provided in the Wall Street Reform and Consumer Protection Act of 2010, commonly known as the Dodd–Frank law.

August–September 2007: The Initial Policy Response

By August 2007, policymakers at the Fed and Treasury recognized (belatedly, critics might say) that impending credit losses from poor lending during the run-up to the housing bubble were not just problems for individual firms or investors but posed a broader threat to the financial system and economy.

The initial response to the manifestation of the crisis in August 2007 relied on conventional tools of monetary policy and moderate regulatory discretion. For example, the Fed made clear in August 2007 that the discount window was available for banks in need, and followed in September with a modest cut in the federal funds interest rate. Treasury officials encouraged efforts by private market participants to avoid fire sales of assets, and shepherded voluntary efforts by mortgage lenders to avoid foreclosures in instances in which the cost of a mortgage modification was less than that of a foreclosure. In Swagel (2009), I discuss these efforts.

With the benefit of hindsight, these policy changes look underwhelming. But at the time, policymakers did not see the need for the extraordinary steps that were eventually taken to respond to the crisis, even setting aside the several legal and political constraints to action that were widely understood to exist. The Treasury could not have gotten the authority to undertake capital injections into private banks in August 2007 even if policymakers had thought this was necessary, and the Fed would have faced a political backlash had it tried under its emergency authority to put into place lending programs for investment banks before Bear Stearns faced failure. Still in late 2007, policymakers did not believe extraordinary action was required, which implies that these legal and political constraints did not bind.

For example, Treasury officials long had been urging financial firms to consider their capital positions, but only the independent bank regulators—notably the Federal Reserve and the Office of the Comptroller of the Currency—had the authority to require banks to fund themselves with more capital rather than by borrowing, or to require that they change their behavior in ways like reducing dividend payments to build capital. Indeed, Timothy Geithner (2014), who as President of the New York Fed was the primary federal regulator for Citigroup, a firm that eventually required extraordinary assistance to survive the crisis, expressed regret in his memoir at not doing more with regard to bank capital. In fairness, given the scope of losses from bad lending and the depth of the subsequent panic, it is not clear that moderate additional amounts of capital would have allowed Citigroup or other firms to avoid the turmoil of 2008. Still, more capital would have helped. Moreover, the Federal Reserve at this time did not regulate the then-investment banks and so could not have required Bear Stearns, Lehman Brothers, Merrill Lynch, Goldman Sachs, or Morgan Stanley to raise more capital—though the US Securities and Exchange Commission could have required this step.

Similarly, the Federal Reserve and the Office of the Comptroller of the Currency (OCC) did not supervise the American International Group (AIG), the insurance company that would require a mammoth bailout. Both regulators did, however, have authority over some of AIG's counterparties in the credit default swaps and securities lending transactions that led to the bailout. With better information and greater foresight, the Fed or OCC might have intervened to limit the accumulation of risk at AIG from the other side (though even here, the Fed and OCC did not supervise investment banks such as Goldman Sachs that were also involved with the AIG transactions).

The failure to respond more strongly to the budding financial crisis in late 2007 reflects many factors, but among them is that policymakers did not fully appreciate the depth of what was to come. Through 2007 and even up to the end of the summer of 2008, mainstream economic forecasts such as from the Congressional Budget Office were for little or no growth in late 2008 and early 2009, but then for a recovery as difficulties in housing and credit markets subsided. Perhaps contributing to the lack of action by financial regulators during the run-up to the crisis is the political reality that it is difficult to rein in financial activity when markets are in an upswing.

The Collapse of Bear Stearns

The response to the collapse of Bear Stearns in March 2008 constituted the first bailout of the financial crisis. Bear Stearns had come to rely on raising short-term liquidity through mechanisms such as repurchase agreements. According to the Securities and Exchange Commission, the firm was meeting its capital requirements in early 2008 (Cox 2008). However, mounting concerns regarding its exposure to real estate–related losses led many investors to stop renewing short-term funding—the functional equivalent of a bank run, as explained in this journal by Brunnermeier (2009). Thus, regulators thought that Bear was solvent, and yet the firm faced collapse within days.

Bear Stearns was not a commercial bank, and so the usual policy responses for a bank facing either liquidity problems or outright failure were not available. As an investment bank, Bear Stearns had neither stable deposit funding backed by Federal Deposit Insurance Corporation (FDIC) deposit insurance nor access to the Fed's discount window for emergency lending support. In addition, if Bear Stearns went broke it would not be resolved like a bank through the time-tested FDIC process discussed by Bovenzi (2015), but instead would go through a standard commercial bankruptcy. Many government policymakers feared that if such a bankruptcy proceeded, Bear's operations would implode as its short-term funding disappeared or through an exodus of clients while the bankruptcy proceeded. In the eyes of policymakers, Bear Stearns was so interconnected with other institutions that its failure could have had systemic consequences as failures on one end of transactions rippled through the financial system. Whether this fear was correct remains a subject of debate. But this belief and the constraint of inadequate legal authority to deal with a failing nonbank financial firm, combined with the sheer rapidity of Bear's collapse, fostered a blunt Fed intervention to facilitate the acquisition of Bear Stearns by JP Morgan Chase.

The Fed turned to its emergency authority under Section 13(3) of the Federal Reserve Act, which at the time said that in "unusual and exigent circumstances," the Federal Reserve could lend to "any individual, partnership, or corporation" so long as the loan was made against adequate collateral in the judgment of the Fed. Note that the requirement was not that the Fed could not actually take losses, but only that the Fed would not expect to take a loss. (As noted below, use of the Fed's emergency lending would later be constrained by the passage of the 2010 Dodd–Frank law.) JP Morgan was willing to buy Bear Stearns, but did not want the transaction to include certain illiquid assets with a notional value of $30 billion. The Fed's solution was to provide financing on these illiquid Bear Stearns assets, with JP Morgan exposed to the first $1 billion of losses.[1] Shareholders of Bear Stearns took large losses, but the bailout ensured that holders of Bear Stearns commercial paper and other obligations were made whole.

The Treasury Department did not have the legal authority to commit taxpayer funds to an intervention—this was granted only in October 2008 with the enactment of the Emergency Economic Stabilization Act that created the Troubled Asset Relief Program. Instead, the Treasury could only provide the Fed with a letter from the Secretary of the Treasury to the Chairman of the Federal Reserve noting that any losses suffered by the Fed would eventually mean smaller transfers of profits from the Fed to the Treasury—that is, the letter offered political cover by acknowledging that the Fed and Treasury were both part of the public balance sheet. In the end, the Fed's loan for the Bear Stearns assets was repaid in full with a $765 million gain from interest payments and increases in the value of the underlying assets. The Fed's action did not require Congressional approval, and the firm's rapid collapse and use of nonrecourse lending to a special purpose vehicle meant that, initially, the transaction was poorly understood in Washington. The backlash against bailouts, however, would build.

Following the collapse of Bear Stearns in March 2008, the Fed put in place the Primary Dealer Credit Facility (PDCF), through which the Fed for the first time since the Great Depression stood ready to lend to the broker-dealer units of investment banks. Though other investment banks such as Lehman Brothers and Merrill Lynch were viewed as vulnerable to large mortgage-related losses, the PDCF was widely seen as ensuring that these firms would not face the sort of funding run that doomed Bear Stearns. In spring 2008, policymakers believed that there would be time instead for these firms to raise additional capital or sell themselves off to stronger institutions while a gradual improvement of the economy would help to stabilize the housing market and asset values with it.

Given the need to rely on the Fed's emergency authority for Bear Stearns, a natural question is whether the Bush administration should have approached Congress in spring 2008 to obtain additional legal power. In March and April 2008, policies discussed inside the Treasury included the possibility of large-scale government purchases of illiquid assets or public capital injections into banks in the event of a broader market crisis. But until such a crisis actually arose, the belief was that lawmakers from both parties would be loath to grant discretionary power to executive branch officials to intervene in private firms and put taxpayer money at risk.

[1] The actual transaction involved a $29 billion Fed loan to a limited liability corporation established by the New York Fed that was combined with $1 billion from JP Morgan to purchase the assets. The corporation was named Maiden Lane LLC; it was named after the street behind the New York Fed main building. If the value of the assets turned out to be less than $30 billion, JP Morgan was exposed to the first $1 billion in losses, after which the Fed took any further losses. In making this loan, the Fed thus asserted that the assets would eventually be worth at least $29 billion. This assumption turned out to be correct, though it was a tenuous assumption at the time.

Indeed, many members of Congress would object to proposals that could be seen as encouraging bailouts by making them more possible.

Others proposed that changes to the bankruptcy code could prove useful for dealing with the crisis, like an idea from Zingales (2008) that the power to convert bondholders into equity shareholders could "immediately make banks solid, by providing a large equity buffer." However, changing the legal constraint preventing such an approach ran into the political constraint. Changes to the bankruptcy code had been enacted with considerable controversy in 2005 after at least seven years of Congressional efforts. Further such changes were simply not possible in a timeframe relevant to dealing with the financial crisis.

The Collapse of Lehman Brothers: Constraints on the Fed and Treasury

The bankruptcy of Lehman Brothers in September 2008 marked the onset of a broad financial panic, leading to questions of why the Federal Reserve did not invoke Section 13(3) to save Lehman. After all, the Fed had made loans for Bear Stearns previously and would make another set of loans within two days of Lehman's failure to prevent the collapse of AIG. The difference between the three situations is that the Fed saw Lehman as insolvent, not only that it was holding illiquid assets, and thus the Fed believed it lacked the legal authority to lend to the firm. This argument raises several questions.

Was the Fed correct in its assessment of Lehman's financial situation? Of course, it was difficult for anyone to determine the valuation of Lehman's assets and liabilities in the fall of 2008, at a time of severe credit market strains under which assets comprised of subprime mortgages were characterized by low liquidity and possibly fire-sale prices. Claims that Lehman's assets might have been worth enough to make the firm solvent or nearly so, such as in Stewart and Eavis (2014), are based on six-year-old recollections and do not match documentary evidence and contemporary accounts. At the time, policymakers and market participants widely believed that Lehman was insolvent, and not merely illiquid, with the firm suffering a capital hole of several tens of billions of dollars (for example, according to Paulson 2010; Geithner 2014). The Fed thus hewed to the law.

Should the Fed have loaned to Lehman Brothers even though central bank officials believed that the firm was insolvent? After all, the law left the evaluation of collateral quality up to the Fed itself and did not provide a mechanism for a third party to object. The law did not prohibit the Fed from taking losses but only from making loans on which it expected to make losses—a vital distinction. This question begins with a recognition that the Fed faced legal constraints and asks whether it should in some cases disregard those constraints. This question might be especially relevant if Fed officials suspected that Lehman's failure would spark a panic and play a role in transforming an economic slowdown into the Great Recession. At the time, however, the Fed and the Treasury did not expect this outcome. While it was widely recognized that Lehman's failure would be challenging for markets because the firm was widely connected to other market participants through derivative contracts and repurchase agreements and because Lehman's failure would call into question the viability of other firms with illiquid assets, the Lehman bankruptcy led to financial panic through two unexpected channels.

First, the Reserve Primary Fund, a large money market fund, had taken a large position in Lehman commercial paper, and the Lehman bankruptcy meant that the fund was forced to "break the buck" by declaring that it could not return investors' money at par. The result was a flight from money market mutual funds as a group. In turn, firms that relied on funding through short-term commercial paper found that it was difficult for them to obtain routine liquidity, because money market mutual funds, which were typically large purchasers of commercial paper, were selling their existing paper to meet redemptions and not buying new issues. The panic in money market funds thus constituted a spillover from the financial sector to the real economy—from Wall Street to Main Street. The Securities and Exchange Commission regulates money market funds, and in principle, could have been aware that the Reserve Fund's exposure to Lehman securities put it at risk, but Lehman paper remained highly rated in the days ahead of the firm's bankruptcy and thus within the scope of allowable assets for money market funds.

Second, the Lehman bankruptcy meant that the assets of many Lehman clients were tied up in London as a result of the UK bankruptcy system, which unlike that in the United States, did not distinguish between the firm's resources and those of its clients for which Lehman was a custodian. This especially affected investment firms such as hedge funds, which in turn sold other assets to generate cash, leading to further downward pressure on asset prices. US policymakers were not prepared for this feature of the British legal system; indeed, the investors whose funds were trapped apparently did not anticipate their dilemma, either.

The panic in money market funds and impact on commercial paper markets was at that time viewed as a grave danger, and Treasury and the Fed both responded by finding ways to use their existing discretionary power. The US Department of the Treasury (2008) used the $50 billion Exchange Stabilization Fund—originally established back in the 1930s to address issues affecting the exchange rate of the US dollar—to set up an insurance program to insure depositors in money market funds. A measure of the panic during that week is that even money market mutual funds that only purchased US government securities bought the Treasury insurance, despite the fact that the federal balance sheet standing behind the insurance was no different than the one standing behind the Treasury securities to be insured. Use of the Exchange Stabilization Fund for this purpose was plausibly legal—after all, a panicked flight from US dollar-denominated securities could be seen as posing a threat to the exchange value of the dollar—but its use in this way was without precedent. Use of the Exchange Stabilization Fund had not been contemplated for dealing with Bear Stearns earlier that year—the rapidity of Bear's collapse and the Fed's response precluded this discussion. In the week following Lehman's collapse when every option was considered, it was clear to

Treasury officials that there would be only one opportunity to use the Exchange Stabilization Fund during the financial crisis because the size of the fund was modest relative to the trillions of dollars that were ultimately guaranteed. This cannon could fire only a single shot. Indeed, Congress was to restrict future use of the Exchange Stabilization Fund as part of the post-crisis reforms, and also limited unexpected uses of government authorities, such as actions by the Federal Deposit Insurance Commission discussed below.

The Fed responded to the related problems in money market funds and commercial paper by developing emergency liquidity programs aimed at these particular markets—steps allowed under the 13(3) emergency authority but extraordinary in that the Fed was offering loans to support an asset class rather than for particular firms. The Money Market Investor Funding Facility provided liquidity to money market mutual funds so that they could avoid fire sales of assets to satisfy the flood of redemptions, and the Commercial Paper Funding Facility effectively served as a buyer of last resort for the new issuance of commercial paper. Together, these programs from the Treasury and Fed were to stanch the redemptions from money market funds. But these programs could only be put in place when the crisis had flared to the point that they were critical—and not beforehand.

While the problems in money markets and commercial paper abated, the panic begun in the week following Lehman's failure continued. Nonetheless, a continuing panic does not suffice to prove that the Fed should have bailed out the firm's funders—this claim requires foresight of the channel through which Lehman's failure affected the economy.

Behind the scenes, top officials from the Treasury and Fed went to extraordinary lengths in seeking to arrange a private solution for Lehman. We will never know for sure because the decision did not have to be taken, but it is possible that the Fed might have been willing to provide some public financing for a transaction if there was a buyer for Lehman that included private capital to absorb potential losses ahead of taxpayers. In the end, and in contrast to the situation with Bear Stearns, no firm was prepared both to absorb at least some of Lehman's losses (perhaps bolstered by Federal Reserve lending) and also actually to continue Lehman's operations. A possible acquisition by the UK firm Barclays would have required a vote by its shareholders at a minimum. It is not clear that British regulators would have allowed the deal in the first place, but they certainly did not allow for the decision to be made rapidly as would be needed for a Fed-assisted transaction.

Having the Fed decide to break its own rules and lend directly to Lehman, despite a lack of sufficient collateral, was not a workable solution. An investment bank dependent on short-term funding implodes rapidly once confidence is lost, and lending by the Fed to Lehman in the absence of a definite plan to sell the firm and have it backed by private capital would probably not have reassured the firm's private sector providers of funding. The end result would have meant that funding from Lehman's private creditors would be replaced by loans from the Fed, leaving American taxpayers exposed to the firm's losses. Moreover, Fed lending to Lehman further would have made market participants expect similar treatment for other teetering firms such as Merrill Lynch (which instead sold itself to Bank of America).

AIG

The Federal Reserve provided some $85 billion in loans to avert the failure of AIG on September 16, 2008, less than two days after not providing support when Lehman Brothers filed for bankruptcy early in the morning on September 15. AIG faced collateral calls from the counterparties to its credit default swaps and securities lending operations. AIG was already pressed to come up with cash and could not meet the additional collateral obligations that followed a September 15 downgrade in its credit rating by Standard & Poor's.

The decision to rescue AIG was driven by two factors. First, the Fed believed that loans to AIG would be adequately secured by a claim against the firm's well-capitalized and profitable global operating subsidiaries. The Fed's judgment that the loan to AIG was made against adequate collateral seems to have been borne out, with the insurer returning to profitability and paying back the government investment with a taxpayer profit. (Taxpayers became involved when Treasury took on the exposure after using resources from the Troubled Asset Relief Program to replace the Fed's lending.)[2] Second, as the world's largest insurance company, AIG was considerably more interconnected with other firms than Lehman, and had substantial consumer- and business-oriented operations so that its failure would have immediate impacts on the real economy.

Legal constraints shaped the way in which the AIG rescue was carried out. The structure of the deal meant that AIG did not declare bankruptcy but instead received loans from the Federal Reserve under a number of onerous conditions. Specifically, the Fed received a one-time fee of 2 percent on its $85 billion loan commitment, an 8.5 percent interest rate on the $85 billion amount, an additional interest rate at the three-month LIBOR yield for cash actually drawn by the company, and rights to 79.9 percent ownership of AIG common stock. AIG presumably accepted the terms at the time because the outcome was better for shareholders and other firm stakeholders than the alternative of bankruptcy. However, these terms are the subject of ongoing litigation as of early 2015.

This intervention by the Fed meant that AIG counterparties such as banks and other counterparties to AIG credit default swaps did not face losses. Shareholders suffered, as was appropriate, but AIG bondholders and others did not. A number of observers have asserted that the Fed should have done more to ensure that at

[2] In this issue, McDonald and Paulson suggest that AIG was perhaps not in fact solvent, and thus that the Fed's decision to lend was based on a mistake in judgment. Placing an accurate valuation on assets and liabilities in September 2008, and distinguishing insolvency from illiquidity, can often involve controversial decisions. As noted above, the key for the Fed was that it believed at the time that its loans to AIG were secured.

least some of the costs and risks of supporting AIG were borne by private investors. Here, legal constraints bound heavily, because no legal authority existed to impose such losses on the counterparties of AIG as a condition of receiving a loan from the Federal Reserve. Indeed, financial regulators in France had forbidden French banks from agreeing to concessions on their claims against AIG. The liabilities of the AIG financial products division were collateralized by the overall AIG balance sheet, so that a refusal by any counterparty to accept a loss would have meant a collapse of the entire firm. Regulators of AIG insurance units across the United States and around the world would have had a fiduciary obligation to grab assets to satisfy policyholders in their local jurisdictions. Counterparties that had already hedged their exposure might actually have ended up worse off had they agreed to concessions than in the event of an AIG default, which meant that they had no incentive to agree to a voluntary haircut. AIG's rapidly deteriorating cash position meant that there was insufficient time to negotiate with its counterparties en masse.[3] The choice was thus to support the firm as a whole or to let it collapse, with the attendant risk of broad negative implications.

Important elements of the Dodd–Frank financial reform legislation in 2010 (officially, the Wall Street Reform and Consumer Protection Act) were put in place in reaction to the constraints highlighted by the Lehman and AIG situations: notably, government officials now have the ability to commit taxpayer funds to prevent the collapse of a systemically important firm that is not a bank, and not just the ability but the obligation to impose losses on equity owners and other counterparties such as bondholders to ensure that the public resources are paid back in full. In future crises, these changes mean that private investors rather than taxpayers will take on the risk and bear the consequences of firms' failures.

TARP and Constraints on Bank Interventions

The Troubled Asset Relief Program (TARP) was proposed on September 18, 2008—the same week as the Lehman collapse and the AIG bailout—and passed into law as part of the Emergency Economic Stabilization Act on October 3, 2008. The TARP provided authority for the Treasury to purchase or guarantee up to $700 billion of troubled assets; in Swagel (2009), I provide details on the development, proposal, and features of the TARP.

[3] One possibility raised by some commentators to sidestep these constraints was for government officials to pressure particular institutions: for example, the Fed and Treasury could have leaned on, say, Goldman Sachs, Merrill Lynch, Bank of America, Citigroup, and Wachovia to accept less than the full amount they were owed by AIG—with those firms specified because they were American institutions that received billions of dollars of collateral posted by AIG (for discussion, see Walsh 2009). Such an action would have treated singled-out firms unequally with others not singled out—including foreign firms with more at stake than these American ones. Fed and Treasury officials brushed off this possibility, making clear both during and after the bailout that there was no alternative in their view but to support AIG as a whole, even with the frustrating implication that all counterparties would be made whole.

The TARP as originally envisioned by Treasury Secretary Paulson was to purchase illiquid mortgage-backed securities to relieve strains in credit markets and provide clarity regarding firms' balance sheets by restarting a process of price discovery for illiquid securities. Implementing the asset purchases involved technical hurdles, including the need to develop a mechanism by which the government would buy the securities and to ensure that the details of the law were followed regarding who could sell to the government.[4] The plan in late September (with work on reverse auctions to purchase assets having begun even before enactment of the legislation) was that small asset purchases could get under way as a proof of concept at the end of 2008 or early 2009. It would take longer for the approach to buy a sizable amount of assets, but there could still be a positive impact sooner than this if the advent of the TARP helped to boost asset values and coax hesitant investors back into the market. Indeed, the mention of the TARP proposal had precipitated a stock market rally.

While the intent of the TARP when it was proposed was to purchase illiquid assets, its switch in focus to capital injections was driven by events and political realities. By the time the TARP was enacted in early October 2008, two more large banks had failed (WAMU and Wachovia). Confidence in the financial system continued to wane, as indicated by measures such as the spread between the low yields on Treasury securities and elevated interest rates for banks to borrow from one another. It became clear to policymakers that a more rapid approach was needed to shore up confidence in the financial system. The switch from asset purchases to capital injections fit within the TARP's legislative language, because shares of banks that originated loans represented troubled assets related to mortgages. Indeed, some members of Congress had urged the Treasury from the start to carry out capital injections rather than asset purchases.

Capital injections could be put in place faster than asset purchases. In addition, each dollar of TARP capacity used for capital injections provided for a greater increase in the loss-absorbing capacity of US banks than a dollar used for asset purchases or guarantees. This is because under the Emergency Economic Stabilization Act of 2008, the purchase or guarantee of an asset such as a mortgage-backed security counted in the same amount against the $700 billion allocated by Congress as the provision of an equal amount of capital directly to financial institutions through the purchase of equity positions. Asset purchases would help cleanse bank balance sheets of illiquid mortgages and contribute to price discovery but would raise firms' net worth only if Treasury intentionally overpaid for assets (which was not the plan) or if asset prices rose following the TARP purchases (a possibility if the implementation of the reverse auctions lifted confidence and thereby improved asset prices).

The Capital Purchase Program (CPP) was announced in a meeting with the chief executive officers of nine large American banks at the Treasury Department

[4] For example, sellers of assets were required to provide the Treasury with warrants on the firm itself, and obey strictures relating to executive compensation.

on October 13—the Columbus Day holiday. The eight institutions ultimately receiving capital injections (after Bank of America's acquisition of Merrill Lynch) together accounted for more than half of both the assets and deposits of the US banking system. The existence of these mega-firms, while giving rise to concerns over institutions that were too big to fail, also made it possible to strengthen a broad swathe of the banking system rapidly. Each firm received public capital equal to 3 percent of its risk-weighted assets, for a total of about $125 billion. The remaining thousands of US banks together would be eligible for another $125 billion in capital.

The use of a broad capital injection, rather than capital provided only to the institutions that needed it most, was driven by policymakers' desire to signal their confidence in the banking system as a whole while providing the resources necessary to reinforce this confidence with loss-bearing capacity. The terms of the capital injections were thus made relatively attractive to ensure broad participation, with banks paying only a 5 percent yield on preferred shares for five years, after which the yield would increase to 9 percent for banks that had not by that time repaid the Treasury. These terms reflected both a legal constraint and a policy purpose: the constraint that it was not possible to require a healthy financial institution to accept a TARP investment, and the policy purpose of encouraging broad participation that would reassure market participants about the overall health of the US financial system. The US approach was in contrast with capital injections in the United Kingdom, which were made on more onerous financial terms, such that relatively strong banks declined to participate.

In 2009, TARP funds were again set to be used to shore up the financial system, serving as the source of public capital backstopping the so-called "stress tests," in which bank balance sheets were evaluated to see whether they could withstand an additional period of financial stress. Banks that lacked the appropriate capital as determined by the stress test would be given a chance to raise additional capital from the private sector after which they would be required by their regulator to accept it from the TARP (on onerous terms meant to induce private capital-raising). Such a mandate was possible for regulators because banks failing the stress tests could be deemed as operating in an unsafe condition. The availability of TARP capital was essential to making the stress tests credible in that public capital was available to be forced on firms that could not (or would not) raise their own in response to the results of the stress test.

Institutional and legal constraints further affected Treasury decisions to provide additional assistance to Citigroup and Bank of America in 2008 and 2009 beyond the initial capital investment of $25 billion for each institution. These two banks (and perhaps others) appeared to be insolvent at points during the crisis, and were to require extraordinary assistance from the TARP, and yet the government propped them up rather than invoking the usual bank resolution authority of the Federal Deposit Insurance Commission. These decisions reflected several factors. First, there was the concern that a government takeover of Citigroup would lead to a renewed flight from other still-fragile banks. Second, while the Federal Deposit Insurance

Corporation had the legal authority to take over each firm's commercial bank, there was little confidence across the government in the agency's ability to run a mega-bank. Taking over a large bank was easier said than actually done—at least before the new powers granted in the Dodd–Frank law. In the end, the shareholders of Citigroup had their ownership stakes substantially diluted by the government investment (including through the conversion of the Treasury preferred stock holdings into common stock), but the firm did not fail. Meanwhile, bondholders and other counterparties avoided losses entirely, which was in some ways less than fully desirable, but did have the positive effect of limiting further financial contagion.

At the same Columbus Day meeting at which the capital injections were announced, the Federal Deposit Insurance Commission introduced the Temporary Liquidity Guarantee Program (TLGP), under which it would insure senior debt issued by banks. The FDIC further extended its deposit insurance to provide an unlimited backstop on business transactional checking accounts that were previously uninsured. The TLGP program was undertaken using the FDIC's emergency authority, which allowed the FDIC to put taxpayer money behind a bank to avoid serious adverse systemic economic or financial effects without the usual requirement to act in a manner that ensured the least cost for taxpayers. Use of this authority required approval by the boards of the FDIC and the Federal Reserve, and the Treasury Secretary was required to consult with the President—all as part of an effort to ensure that the authority was not used lightly. Introduced in the Federal Deposit Insurance Corporation Improvement Act of 1991, the systemic risk exception had not been used until earlier in September 2008, when the FDIC sought to use it as part of the transaction by which Citigroup was to buy the failing Wachovia bank (in the end, Wells Fargo instead purchased Wachovia without government assistance). The Dodd–Frank legislation was later to prohibit a repeat of the TLGP without explicit Congressional approval.

Veronesi and Zingales (2010) calculate that the guarantees from the Federal Deposit Insurance Corporation account for most of the benefits in terms of stabilization of the financial system. This raises the question of whether the TARP capital injections could have been avoided in favor of just the FDIC guarantees along with the expansions of Fed liquidity, such as for commercial paper and eventually securitized assets under the Term Asset-Backed Securities Loan Facility (TALF), and the Fed purchases of Treasury and mortgage-backed securities under its quantitative easing policies. After all, the FDIC and Fed actions were undertaken with existing emergency powers and did not require Congressional action. Indeed, one can argue that the TARP legislative process itself may have contributed to increased uncertainty in late September 2008 that could have been avoided by limiting action to the Fed and FDIC.

However, this scenario of proceeding without something like the TARP program was infeasible. The guarantees from the Temporary Liquidity Guarantee Program of the Federal Deposit Insurance Corporation would never have been put in place without the existence of the TARP program. While all sources of taxpayer funds are on the same balance sheet, the FDIC in practice acts as if this is not the case, seeking to protect its deposit insurance fund to avoid having to utilize the statutory authority to borrow from the Treasury.[5] Without the advent of the TARP and its use for capital injections, the FDIC would have feared that its expanded bank guarantees would create too high a risk of needing to borrow from the Treasury, and thus the FDIC have not agreed to put in place the TLGP.

Another suggestion is that the capital injections should have been put into place sooner—that such action had been part of other financial rescues and the Treasury should have learned this lesson from other nations such as Sweden. The difficulty is that the Emergency Economic Stabilization Act legislation that allowed the eventual capital injections would not have been enacted if the proposal presented to Congress were for the US government to purchase $700 billion stakes in private banks. This was a hard political constraint. The legal constraints preventing the TARP capital injections—the response that was ultimately essential to resolving the crisis—could only be addressed when the crisis had become serious enough that political constraints dropped aside. And this was the case only when the use of pre-existing emergency authority by the Fed and FDIC was not enough to arrest the mounting financial sector panic.

Conclusion: Implications for the Next Crisis

What constraints will policymakers and regulators face when the next financial crisis arrives? It seems safe to conclude, based on political considerations, that there will not soon be another Troubled Asset Relief Program, with its broad grant of authority for the government to put taxpayer money into the financial system. Attacks on the bank bailouts in particular have become a staple of political campaigns. Moreover, some emergency actions taken during the crisis are no longer available to policymakers as a result of provisions in the 2010 Dodd–Frank financial reform bill. The Treasury is no longer permitted to use the Exchange Stabilization Fund to guarantee money markets. The Federal Deposit Insurance Corporation must now obtain Congressional approval to provide broad debt guarantees. The Federal Reserve can no longer make emergency loans to individual nonbank institutions but must instead devise broad-based programs.

At the same time, the Dodd–Frank law provided important new powers for government regulators to respond to a future financial crisis. Title II of the Dodd–Frank law creates a nonbank resolution authority under which the government can put taxpayer funds into a failing institution to prevent a collapse. Government officials are required to recoup taxpayer funds by imposing losses on shareholders, bondholders, or other counterparties of the failing firm, and

[5] The desire of the FDIC to avoid borrowing from the Treasury could be seen in the September 2009 action to have banks pre-pay for future deposit insurance premiums as a way of adding resources to the insurance fund (Labaton 2009), even though this imposed a drag on bank resources at the same time as banks were being urged to expand their lending to support the recovery.

ultimately through assessments on other financial sector participants if needed. The Federal Deposit Insurance Corporation is still developing the tools for such an intervention. However, the broad approach is similar to that taken with AIG, in which taxpayer funds go to the parent company and stabilize the firm as a whole. Bovenzi, Guynn, and Jackson (2013) discuss this Title II authority, including the relationship with the bankruptcy code.

Other legal and institutional changes also address weaknesses highlighted by the financial crisis. The Financial Stability Oversight Council (FSOC) was put in place by the Dodd–Frank legislation to avoid the situation with AIG, in which risks developed in a lightly regulated part of the financial system. The FSOC is meant to give all regulators, but especially the Fed, the affirmative duty to pay attention to risks anywhere in the financial system, while the Office of Financial Research established under Dodd–Frank is meant to contribute to this effort as well. These institutional innovations so far do not appear to have had much effect, though it is too soon to know the eventual outcome.

Banks in the wake of the financial crisis are funded with considerably more capital than previously, and are required to ensure that they have stable access to increased sources of liquidity. Many derivative transactions are required to take place on exchanges and through clearinghouses, providing financial regulators with greater ability to assess risks that were previously opaque. A Consumer Financial Protection Bureau was created to address problems highlighted by the crisis, including a lack of clarity or disclosure in financial products.

Given these new legal authorities, it seems clear that the policy response to a future crisis would face different constraints and thus would unfold in a different way. It could be that the increased and altered ability of government officials to intervene during a time of crisis leads to unexpected negative consequences. Bondholders in the last crisis assumed that some banks were too big to fail—and they were right—and thus counted on an intervention that made them whole. With the Title II resolution authority, however, the government can seize a large troubled firm and impose losses on bondholders while maintaining the firm's operations to avoid a broader financial market fallout. In the future then, it could be that systemically important firms subject to the Title II resolution authority will find that their funding dries up rapidly in the face of difficulties, as bondholders and sources of liquidity pull away to avoid the losses. In other words, the ability of policymakers to seize a large financial firm could cause such firms to lose their funding more quickly, thereby making this kind of intervention more likely. It will be hard to know until the next crisis.

In the meantime, I prefer to think of the glass as half full. Political constraints meant that the essential step of the TARP was proposed only when the financial crisis was severe enough to make possible Congressional action to avoid economic meltdown. While there will not be another TARP, the post-crisis reforms have given policymakers certain essential authorities that did not exist in 2007 to 2009—the ability to stabilize a troubled but systemically important firm while imposing losses on private market participants. Indeed, the understanding that such losses are required in the future should affect markets today; potential lenders to large banks will likely reassess the returns they require knowing that by law they must take losses in a future crisis rather than receiving a bailout. In sum, an understanding of the political and legal constraints that affected the policy response in 2007 to 2009 has the potential to make the future response yet more effective and the next crisis less damaging.

■ *I thank David Autor, Chang-Tai Hsieh, Randy Kroszner, Ann Norman, Hank Paulson, and Timothy Taylor for comments that improved the paper.*

References

Bovenzi, John F. 2015. *Inside the FDIC: Thirty Years of Bank Failures, Bailouts, and Regulatory Battles.* Wiley.

Bovenzi, John F., Randall D. Guynn, and Thomas H. Jackson. 2013. *Too Big to Fail: The Path to a Solution.* A Report of the Failure Resolution Task Force of the Financial Regulatory Reform Initiative of the Bipartisan Policy Center, May. Bipartisan Policy Center.

Broda, Christian, and Jonathan A. Parker. 2014. "The Economic Stimulus Payments of 2008 and the Aggregate Demand for Consumption." NBER Working Paper 20122, May.

Brunnermeier, Markus K. 2009. "Deciphering the Liquidity and Credit Crunch 2007–2008." *Journal of Economic Perspectives* 23(1): 77–100.

Cox, Christopher. 2008. "Chairman Cox Letter to Basel Committee in Support of New Guidance on Liquidity Management." Securities and Exchange Commission, March 20. http://www.sec.gov/news/press/2008/2008-48.htm.

Geithner, Timothy F. 2014. *Stress Test: Reflections on Financial Crises.* Crown Publishers.

Labaton, Stephen. 2009. "Banks to Prepay Assessments to Rescue F.D.I.C." *New York Times*, September 29. http://www.nytimes.com/2009/09/30/business/economy/30regulate.html.

Paulson, Henry M. Jr. 2010. *On the Brink: Inside the Race to Stop the Collapse of the Global Financial System.* Business Plus, Grand Central Publishing.

Stewart, James B., and Peter Eavis. 2014. "Revisiting the Lehman Bailout That Never Was." *New York Times*, September 29. http://www.nytimes.com/2014/09/30/business/revisiting-the-lehman-brothers-bailout-that-never-was.html.

Swagel, Phillip. 2009. "The Financial Crisis: An Inside View." *Brookings Papers on Economic Activity*, Spring, pp. 1–63.

US Department of the Treasury. 2008. "Treasury Announces Guaranty Program for Money Market Funds." Press Release, September 19. http://www.treasury.gov/press-center/press-releases/Pages/hp1147.aspx.

Veronesi, Pietro, and Luigi Zingales. 2010. "Paulson's Gift." *Journal of Financial Economics* 97(3): 339–68.

Walsh, Mary Williams. 2009. "A.I.G. Lists Banks It Paid with U.S. Bailout Funds." *New York Times*, March 15. http://www.nytimes.com/2009/03/16/business/16rescue.html.

Zingales, Luigi. 2008. "Plan B." *The Economist's Voice* 5(6): Article 4.

Journal of Economic Perspectives—Volume 29, Number 2—Spring 2015—Pages 123–150

After publication, the author discovered some coding errors, which do not substantially change anything in the analysis. A corrected version is now posted online with the supplementary materials.

Understanding the Increase in Disability Insurance Benefit Receipt in the United States[†]

Jeffrey B. Liebman

The share of working-age Americans receiving disability benefits from the federal Disability Insurance (DI) program has increased significantly in recent decades, from 2.2 percent in the late 1970s to 3.6 percent in the years immediately preceding the 2007–2009 recession and 4.6 percent in 2013.

Some experts have interpreted the increase as evidence of a need for significant reform. In this journal, Autor and Duggan (2006) describe the growth in the disability insurance rolls as "a fiscal crisis unfolding," report that "abuse [has] reached unsustainable levels," and conclude that "the DI screening process is effectively broken." In their view, changes in program rules enacted in 1984 made it easier for applicants to receive benefits for hard-to-verify impairments like back pain and depression. In conjunction with labor market developments that increased the incentive for low-wage workers to apply for benefits, these new program rules led to an increase in disability receipt.

Other experts attribute most of the increase in beneficiaries to baby boomers reaching their peak disability-claiming years and to increased labor force participation by women, which has made more women eligible to claim disability benefits (Reno 2011). Under this interpretation, disability enrollment rates and spending are unlikely to rise much further, because these demographic trends have largely run their course. Indeed, both the Social Security Administration actuaries (OASDI

■ *Jeffrey B. Liebman is Malcolm Wiener Professor of Public Policy, John F. Kennedy School of Government, Harvard University, Cambridge, Massachusetts. He is also Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is jeffrey_liebman@harvard.edu.*

[†] To access the Data Appendix, visit http://dx.doi.org/10.1257/jep.29.2.123

doi=10.1257/jep.29.2.123

Board of Trustees 2014) and the Congressional Budget Office (CBO 2012) project that spending on Disability Insurance will fall as a share of GDP in the coming decade as baby boomers convert from DI benefits to retirement benefits and are replaced in the peak disability-receiving ages by smaller cohorts.

With the federal Disability Insurance Trust Fund currently projected to be depleted in 2016, Congressional action of some sort is likely to occur within the next several years. It is therefore a good time to sort out the competing explanations for the increase in disability benefit receipt and to review some of the ideas that economists have put forth for reforming US disability programs.

The resolution of the competing explanations is a tale of two time periods. During the 1980s, policy changes caused receipt of Disability Insurance benefits first to plummet and then to rebound. In this period, the overwhelming majority of the change in disability benefit receipt came from changes in "incidence rates" (of new awards among the insured not already receiving benefits), though increased eligibility for benefits among women also played a role. Since the early 1990s, incidence rates among men, adjusted for the population age distribution and the business cycle, have been steady, while those for women have been gradually approaching those of men. In this period, population aging and increased eligibility among women account for two-thirds of the increase in DI benefit receipt, rising incidence among women accounts for one-quarter, and declining mortality rates account for one-sixth.

While adjusted incidence rates have mostly leveled off, there has been a change in the composition of DI recipients, with more recipients claiming benefits for hard-to-verify impairments and with the program playing an increasingly important role in providing income for low-skilled workers whose economic prospects have stagnated. Thus, the case for DI reform is not primarily a fiscal one—up until the 2007–2009 recession, spending on the program as a share of GDP had increased by only 0.13 percent of GDP over 30 years. Instead, it is about re-optimizing the program in light of the changing characteristics of the beneficiary population.

The US Disability Insurance System

The Social Security Administration projects that one-quarter of today's 20 year-olds will become disabled and receive benefits from the Disability Insurance program for some period of time before reaching age 67 (Social Security Administration 2014b). Thus, disability is a major economic risk—typically combining less ability to earn income with higher health-related costs—against which people should desire insurance. In theory, one could imagine a private system in which workers voluntarily purchase disability insurance throughout their careers; in practice, many if not most workers would fail to purchase such insurance. Moreover, the challenge of regulating a private disability insurance market to minimize both adverse selection and litigation over eligibility for insurance payments would be significant (Mashaw 1983). Thus, there is a rationale for a compulsory disability insurance system based on the myopia of consumers and the problems that would be faced by private insurance markets in this area, just as there is for Social Security retirement benefits (Feldstein and Liebman 2002).

There are two main federal disability benefit programs in the United States that assist individuals with severe impairments. Social Security Disability Insurance (DI), the focus of this paper, is a contributory social insurance program that replaces lost wages of people with significant work histories. Supplemental Security Income (SSI) is a means-tested program that provides benefits to low-income disabled, blind, or aged people regardless of work history; SSI spending on disabled individuals accounts for approximately 80 percent of all SSI benefits.[1] In addition to cash benefits, these programs confer eligibility for government-provided health insurance—Medicare in the case of SSDI and Medicaid in the case of SSI.

The Social Security Act defines "disability" as the inability to engage in substantial gainful activity because of a medically determinable physical or mental impairment that is expected to last at least 12 months or result in death. To operationalize this definition, the Social Security Administration uses a five-step sequential process. The first two steps disqualify applicants who are currently earning above the substantial gainful activity limit ($1,090 per month in 2015) or who do not have a severe impairment or combination of impairments that is expected to last 12 months or result in death. The third stage compares the applicant's impairment to a listing of impairments, for each major body system, that are considered severe enough to prevent an individual from gainful activity. For example, someone with aggressive lymphoma will meet the listing level of disability and automatically qualify for benefits. For an applicant whose impairments do not automatically meet the listings, the SSA moves to the fourth stage, which involves assessing the person's residual functional capacity and considering whether the individual's impairments prevent the person from doing his or her past work. If so, the individual then moves to the fifth stage of the process, where the SSA considers the applicant's age, education, and work experience—known as the "vocational grids"—and decides whether the person's residual functioning capacity together with his or her place in the vocational grids prevents the applicant from doing other work that exists in the economy. For example, a 50 year-old applicant who is restricted by his impairment to do no more than sedentary work, has no transferable skills to do other work, and has a high school education or less will be found to be disabled, whereas a 50 year-old with more education and with transferrable skills to do other work would not be found to be disabled.

These standards are applied in three main stages. Disability examiners at state Disability Determination Service (DDS) offices make an initial determination. An applicant who is denied can appeal to be reconsidered by another disability examiner at the same DDS office. If the applicant is denied a second time, the applicant can appeal for a hearing before an Administrative Law Judge (ALJ).

Determining whether an individual is eligible to receive disability benefits is much more complicated and requires significantly more administrative judgment

[1] There are also more narrowly targeted disability benefit programs for veterans, railroad employees, and federal civilian employees.

than the determination of eligibility for other large social insurance programs like Social Security retirement benefits, where eligibility is triggered by reaching the eligibility age, or Unemployment Insurance benefits, where eligibility is triggered by an involuntary job separation. The administrative complexity of the disability system, combined with limited agency resources, has resulted in long delays in determining eligibility and in disability allowance rates that vary significantly depending on the office, examiner, or Administrative Law Judge to which a case is assigned (Rupp 2012; Maestas, Mullen, and Strand 2013).

Approximately 65 percent of Disability Insurance applications are resolved at the initial determination stage, while 35 percent are appealed. Most of those who appeal eventually have a hearing before an Administrative Law Judge. In 2008, out of every 1,000 initial applications, 366 were allowed at the initial determination, and 283 of those who were denied did not appeal. Of the 351 applicants who appealed (a 55 percent appeal rate among those who were initially denied), 215 were ultimately allowed at the reconsideration or appeals level. Overall, 58 percent of applicants were allowed benefits, 28 percent were denied without appeal, and another 14 percent were denied after appeal (Social Security Administration 2014a).

For applications that are resolved at the initial stage, average wait times for a determination are generally between 100 and 120 days. However, for those receiving an Administrative Law Judge (ALJ) hearing, the delays are often quite long. When the backlogs were at their worst in August 2008, applicants had to wait 532 days on average for an ALJ hearing, in addition to the time spent waiting for an initial decision and a reconsideration. Management focus and additional resources for ALJs reduced the average wait times to 340 days in October 2011, but recent budget cutbacks and the surge in applications during the recession caused wait times for ALJ hearings to climb again—to 396 days at the end of 2013.

Benefit levels for Disability Insurance are determined by the same benefit formula used for Social Security benefits: that is, benefits (in 2015) replace 90 percent of the first $826 dollars of prior monthly earnings, 32 percent of monthly earnings between $826 and $4,980, and 15 percent of monthly earnings above $4,980. The calculation of prior earnings for disability benefits is based on a worker's average indexed earnings in the years before the person became disabled.[2] In addition, DI benefits are not reduced when claimed earlier in life, whereas approximately 80 percent of Social Security retirement beneficiaries claim benefits before the "full benefit age" and have their benefits reduced accordingly. The average monthly benefit for a disabled worker is $1,146 and the interquartile range on the share of pre-tax lifetime indexed earnings that is replaced by these benefits extends from approximately 45 percent to 80 percent (Muller 2008). Accounting for taxes and

[2] In calculating the average indexed earnings, only the highest *y* years of indexed earnings count, where *y* is the number of years between the year the person turned age 22 and the year the person became disabled, minus between two and five "dropout" years (those with greater elapsed time between age 22 and becoming disabled are entitled to more dropout years).

the Medicare benefits associated with DI receipt would increase these replacement rates (Autor and Duggan 2006).

Several major legislative changes in recent decades have altered disability eligibility criteria and how the criteria are administered. During the 1970s, spending on Disability Insurance benefits increased rapidly as Congress raised Social Security benefit levels and made an error in setting the inflation indexing formula that was particularly significant in that high-inflation era. During this period, the median DI replacement rate increased substantially, creating an increased incentive for workers to apply for DI benefits, and administrative cutbacks reduced the review of state disability awards (Kearney 2005/2006). Concern about program costs led to a tightening of medical eligibility standards and to the Social Security Amendments of 1980. Among other provisions, these amendments required the Social Security Administration to conduct Continuing Disability Reviews to reevaluate beneficiary eligibility every three years except for those beneficiaries whose disability was expected to be permanent.

In the early 1980s, these Continuing Disability Reviews terminated benefits for 490,000 beneficiaries, with 200,000 of the terminations reversed upon appeal (Kearney 2005/2006). These terminations brought a strong political backlash. By 1984, 17 governors had suspended the reviews in their states. One reason that the terminations were perceived as unfair is that medical standards had been tightened, and the reviews applied the new standards—causing beneficiaries to be removed from eligibility even though their medical conditions had not improved. The fact that the bulk of the terminations occurred during a deep recession added to their unpopularity.

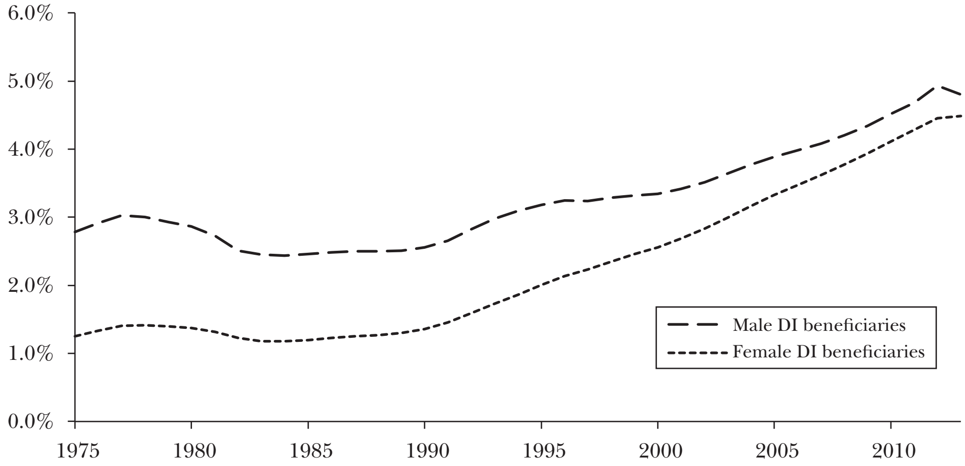
Congress reacted with the Social Security Amendments of 1984, which restricted the circumstances under which disability benefits could be terminated. Under the new law, benefits could be terminated only if the beneficiary experienced a medical improvement or if the government could demonstrate that the initial determination was in error. The Amendments also required the Social Security Administration to develop new standards for individuals with mental disorders, to evaluate pain as part of the disability determination process, to consider the effects of multiple nonsevere impairments in determining disability, and to place greater emphasis on evidence from the applicant's treating physician in the disability determination process. In the wake of these reforms, the disability rolls expanded, reversing the trend of the preceding years. Since then, the basic framework for Disability Insurance has remained much the same.[3]

The Rise and Shifting Composition of Disability Enrollment

The share of working-age Americans receiving disability benefits from the federal Disability Insurance (DI) program is shown in Figure 1 for the years 1975 to

[3] One other significant change occurred in 1996, when legislation was enacted that made individuals ineligible for benefits if drug addiction or alcoholism played a significant role in their disability.

Figure 1
Percent of Working-Age (20–64) Population Receiving Disability Insurance (DI) Benefits, 1975–2013



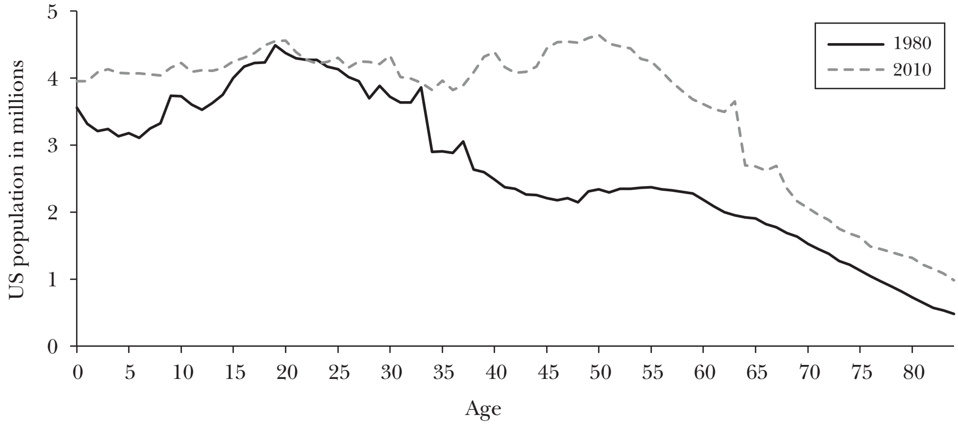
Sources: Social Security Administration, Office of the Chief Actuary; and author's calculations.

2013. The fraction of men receiving DI increased from 3.0 percent in the late 1970s to 3.9 percent in the years immediately preceding the 2007–2009 recession and nearly 5 percent in 2013. Among women, DI receipt increased from 1.4 percent in the late 1970s to 3.6 percent in 2007 and 4.5 percent in 2013.

Over the same period during which these increases in disability enrollment rates were occurring, major demographic changes were occurring as well. As the baby boom generation born after World War II has moved through the work force, it first increased the number of young workers, who are less likely to be disabled, and then in recent years has swelled the number of workers in their late 50s and early 60s, who are the group most likely to be receiving disability benefits. Figure 2 shows the number of Americans of each age in 1980 and 2010. In 1980, there were approximately 23 million individuals between the ages of 50 and 59. By 2010, there were over 42 million. Figure 2 also shows that the cohorts behind the baby boomers are somewhat smaller, partially explaining why the Social Security Administration is predicting spending on Disability Insurance to decline over the coming decade. Americans who are between the ages of 50 and 64 are four and one-half times as likely as those between the ages of 20 and 49 to be receiving Disability Insurance benefits (that is, about 9 percent for the older age group compared to 2 percent for the younger age group). Thus, an increase in the share of the working-age population that is at the peak disability-claiming ages can result in significant changes in overall disability enrollment rates.

The other relevant demographic change occurring over this time period is the increase in the fraction of women with significant labor market experience. To be eligible for Disability Insurance benefits, a worker generally needs to have worked

Figure 2
US Population by Age, 1980 and 2010



Sources: US Census Intercensal Population Estimates (accessed via NBER.org) and author's calculations.

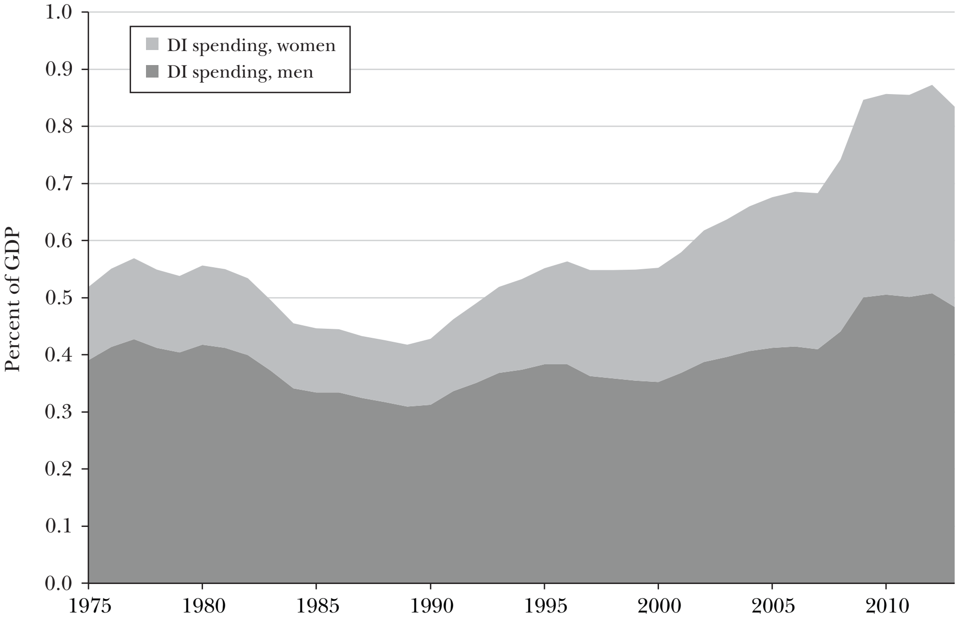
in five of the past ten years.[4] As women entered the labor force in large numbers, the fraction of women ages 50 to 64 "covered" by Disability Insurance—that is, eligible by their work history to receive disability benefits—rose from 46 percent to 72 percent between 1980 and 2007.

The increase in spending on Disability Insurance has not been as great as the increase in enrollment rates. Figure 3 shows spending on DI benefits from 1975 to 2013. DI benefits for men were 0.4 percent of GDP in the late 1970s and were also 0.4 percent of GDP in the years leading up to the 2007–2009 recession. In between, spending fluctuated with the business cycle and legislative changes. For women, spending increased from 0.14 percent of GDP in the late 1970s to 0.27 percent of GDP in 2007, with spending as a share of GDP increasing steadily from 1989 onward. Overall spending on DI benefits increased by 0.13 percent of GDP between the late 1970s and the years preceding the 2007–2009 recession: specifically, from 0.55 to 0.68 of GDP. In comparison, spending on Medicare and Medicaid increased by 3.2 percent of GDP over the same time period, increasing every year by approximately the same percent of GDP as DI spending increased over the entire 30 years.

The reason that spending relative to GDP has risen by only 22 percent when enrollment rates have risen by nearly 80 percent is that benefits have not kept up with productivity growth. Average benefits from Disability Insurance have fallen relative to per worker GDP because these benefits depend on the prior earnings levels of recipients, and there has been: 1) a decline in the worker compensation share of GDP;

[4] To be eligible for disability benefits, a worker generally needs to have earned 40 work credits, 20 of which need to be earned in the last 10 years ending with the year the worker became disabled. In 2015, workers receive one credit for each $1,220 of annual earnings with a maximum of four credits earned in any calendar year. The credit requirements are reduced for workers who become disabled at younger ages.

Figure 3
Spending on Disability Insurance (DI) Benefits, 1975–2013
(as percent of GDP)



Sources: Annual Statistical Report on the Social Security Disability Insurance Program (Social Security Administration 2011); Annual Statistical Supplement to the Social Security Bulletin (Social Security Administration 2012, Table 7.A5; 2013, Table 4.A2); 2013 Economic Report of the President; and author's calculations.
Notes: Allocation between males and females is based on December data from each year. Benefits for spouses and dependents are allocated between the sexes in proportion to worker benefits. The male–female split in DI benefits is interpolated between 1975 and 1980 and between 1980 and 1985.

2) an increase in health benefits as a share of compensation (and a decline in the earnings share); 3) a decline in the ratio of earnings "covered" by Disability Insurance to total earnings resulting from a rise in earnings inequality; and 4) a shift in the earnings distribution of the DI-claiming population toward those with lower earnings.[5]

[5] Specifically, spending relative to GDP can be decomposed into average benefits relative to per worker GDP and the enrollment rate, where the growth in per worker GDP can be thought of as analogous to productivity growth:

$$\frac{Spending_t}{GDP_t} = \frac{\overline{Benefits}_t}{GDP_t} \times Recipients_t = \frac{\overline{Benefits}_t}{GDP_t/WAPop_t} \times \frac{Recipients_t}{WAPop_t}.$$

For example, from 1977 to 2006, DI recipients as a share of the working-age population (WAP) grew by 68 percent, while average benefits relative to GDP per WAP fell by 26 percent. Spending relative to GDP rose by 24 percent (1.68 × 0.74 = 1.24). See Liebman (2014) for further details.

Spending on Medicare benefits provided to recipients of Disability Insurance is about two-thirds as large as spending on cash benefits. It has also risen faster than the disability enrollment rate—from 0.12 percent of GDP in the late 1970s to 0.39 percent of GDP in the pre-recession years—because health care spending per beneficiary has historically risen faster than GDP. That said, given the expansions of Medicaid eligibility and subsidies for insurance purchase enacted as part of the Patient Protection and Affordable Care Act of 2010, many DI recipients would today be receiving free or heavily subsidized health insurance even if they were not receiving disability benefits.

Decomposing the Rise in Disability Enrollment

The rise in disability enrollment has resulted from a mixture of factors: major demographic trends, changes in program rules and implementation, and evolving economic conditions. But how much of the change in disability enrollments can be attributed to each factor?

The methodology I use to answer this question is straightforward. I model the number of people of age *a* who are receiving benefits—"in current payment" (ICP)—in year *t*. The number of people in current payment increases with new disability awards and declines with terminations. New awards are the product of the incidence rate and the number of exposed individuals (the insured population minus those already receiving benefits). Terminations come through death or recovery.[6] "Recovery" is often an involuntary removal from benefit status that occurs when the Social Security Administration performs a Continuing Disability Review and determines that benefits were awarded in error or that the individual's health status has improved. In the model, *a* represents single years of age from 20 to 64.

$$\begin{aligned} ICP_{at} &= ICP_{(a-1,\, t-1)} + \textit{new awards}_{at} - \textit{terminations}_{at} \\ \textit{new awards}_{at} &= \textit{incidence}_{at}((\textit{population}_{at} * \%\textit{insured}_{at}) - ICP_{(a-1,\, t-1)}) \\ \textit{terminations}_{at} &= (\textit{death rate}_{at} + \textit{recovery rate}_{at}) * ICP_{(a-1,\, t-1)}\,. \end{aligned}$$

The model can be used to examine counterfactual scenarios in which one or more parameters are held constant so as to analyze the share of the change over time that can be attributed to changes in the age distribution of the population, the insured rate, the incidence rate, the death rate, and the recovery rate.

The data for the model come from the Office of the Chief Actuary at the Social Security Administration. The raw data contain all of the elements in these three equations, aggregated to five-year age ranges. I interpolate linearly between

[6] At the Social Security full benefit age, terminations can also occur from individuals transitioning to retirement benefits. The results in this paper are limited to individuals 64 and younger. This avoids complications associated with the on-going increase in the Social Security full retirement benefit age from 65 to 67.

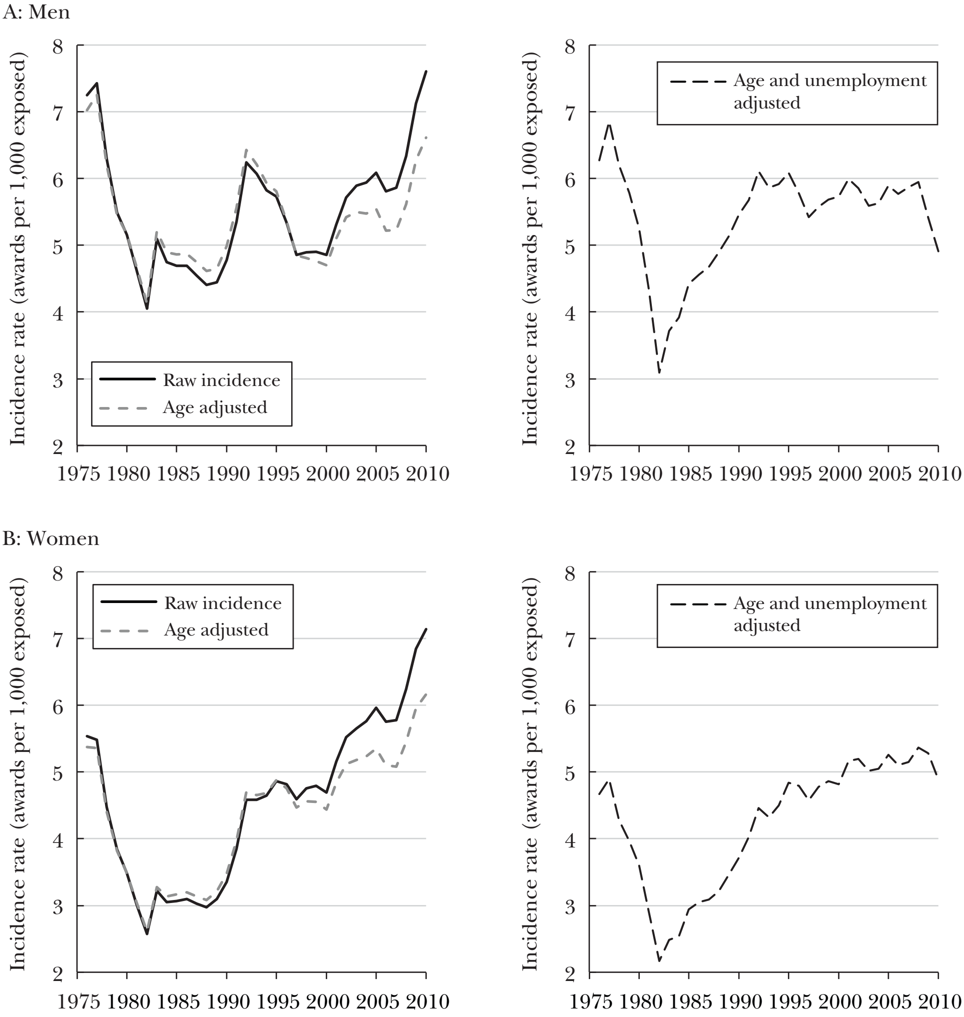
the midpoints of the age ranges to produce data at the level of individual years of age. The model successfully captures the evolution of the number of individuals in current payment over time.

There are several decisions to make in choosing which counterfactual scenarios to analyze. First, which time period to analyze? As discussed above, Disability Insurance enrollment plummeted in the early 1980s before rebounding in the second half of the 1980s. An analysis that takes 1985 as the base year will attribute much more of the change over time in enrollment to incidence than one that takes 1980 or 1990 as the base. In this analysis, I focus primarily on the 1985–2007 period in order to inform discussions about how enrollment rates have evolved since the 1984 legislative reforms. However, I also present results for 1977–2007 and for the 1977–1985, 1985–1993, and 1993–2007 subperiods to highlight the fact that different factors are responsible for a different share of the rise in DI enrollment in different time periods. I stop the simulations in 2007 because my focus is on the long-run program trends rather than the particular impact of the deep 2007–2009 recession. The DI enrollment rate increased by about 1 percentage point during the recession. Cutler, Meara, and Richards-Shubik (2012) find that the recession-induced increase in DI claiming was similar to that in prior recessions. My own estimates described in Liebman (2014) indicate that the rise in DI claiming during the 2007–2009 recession was somewhat lower than would have been predicted by the previous relationship between unemployment and incidence. It is possible that the availability of extended unemployment insurance benefits in the recent recession prevented some DI claiming (Rutledge 2011). However, Mueller, Rothstein, and von Wachter (2013) find "no indication that expiration of UI benefits causes DI applications."

A second analytic choice is how to stack the various parameters. The impact of rising incidence on the Disability Insurance enrollment rate will be greater if demographic changes such as population aging and increased female labor force participation have resulted in more insured individuals in the age range in which disability receipt is most common. Similarly, the impact of demographic changes on the enrollment rate will be larger if incidence is higher. To address this issue, I attribute to incidence the increase in enrollment rates that would have occurred absent population aging and changing insured rates. Separately, I estimate the effect of population aging and changing insured rates under a counterfactual scenario in which incidence rates remained constant. The sum of these separate estimates is smaller than the total effect when all three factors are held constant together. I classify the difference between the separate effects and the total effect as "interaction effects." For simplicity, I stack the two quantitatively less-important factors—mortality rates and recovery rates—at the end of the analysis and do not estimate separate interaction effects for them. This results in my methodology attributing somewhat less impact to declining mortality rates than would occur if I stacked that parameter earlier in the analysis.

A final analytic choice is which year to treat as the base year for each parameter. It is not possible to choose a single year like 1985 as the base year for all of the factors, because some of them exhibited extreme values immediately after the 1984 reforms. Most of the choices are straightforward, and I describe them as I present

Figure 4
Incidence Rates for Men and Women, Ages 20–64



Sources: Social Security Administration, Office of the Chief Actuary; and author's calculations.
Notes: In Figure 4A, the graph on the left shows the actual incidence rate for men, along with an age-adjusted rate that holds the age distribution of the population constant at its 1980 level, while the graph on the right shows the predicted male age-adjusted incidence rate, under the counterfactual assumption that unemployment rates were constant at the 1975–2010 mean value of 6.3 percent for the entire period. Figure 4B presents a parallel analysis for women.

the results below. However, the choice of a base for the incidence rate requires more discussion because applications for disability benefits vary considerably over the business cycle (Autor and Duggan 2003).

The top left panel of Figure 4A shows the actual incidence rate for men, along with an age-adjusted rate that holds the age distribution of the population constant at its 1980 level. Four patterns are evident. First, incidence rates are highly cyclical, rising sharply in response to the 1990–1991, 2001, and 2007–2009 recessions. Second, incidence plummeted after the late 1970s and early 1980s reforms that tightened eligibility and increased the number of continuing disability reviews (CDRs), before rebounding after 1982 and particularly after the 1984 legislation that altered eligibility rules and standards for CDRs. Third, since 1985 there appears to be an upward trend in the actual incidence rate. Fourth, the post-1985 upward trend is less steep in the age-adjusted incidence rate, but it is hard to isolate the trend visually given the large business-cycle-related fluctuations that are occurring throughout this period.

To isolate the underlying time pattern of incidence from business cycle fluctuations, the top right panel of Figure 4A shows the predicted male age-adjusted incidence rate, under the counterfactual assumption that unemployment rates were constant at the 1975–2010 mean value of 6.3 percent for the entire period. These predictions use coefficients obtained from separately regressing the annual incidence rate for each of nine five-year age ranges on the contemporaneous unemployment rate and a one-year lag in the unemployment rate, using a methodology similar to that of the Social Security Technical Advisory Panel (2011).[7] The unemployment-adjusted series reveals a much more pronounced increase in male incidence in the years following the 1984 legislation—a pattern that was obscured in the top left panel by the high unemployment rates of the 1980s, which inflated disability incidence rates relative to what they would have been with more typical unemployment rates. In addition, the unemployment-adjusted series indicates that there has been no increase in incidence among men since the early 1990s.

Figure 4B repeats this analysis for women. The unemployment-adjusted series similarly exhibits a steep rise in incidence after 1984. It also shows that, different from men, incidence has continued to rise for women since the early 1990s, but at a slower rate than during the 1980s. Indeed, incidence for women is now approaching the level for men.

Next we will look at some counterfactual simulations to interpret the impact of various factors on the percentage of the working-age population receiving disability insurance. The analysis of Figure 4 demonstrated that to interpret the impact of incidence correctly, one needs to adjust for the business cycle. Simply using the 1985 incidence rate as the base year for simulations would lead one to understate the contribution of rising incidence rates to the increase in the disability insurance beneficiary ratio because, as just noted, 1985 was a high unemployment year. So to begin, I first modify the actual beneficiary to working-age population ratio to provide an alternative series that projects the path that the ratio would have taken if the unemployment rate had remained steady at its average value for the entire time period analyzed for the simulation. This is done by allowing all of the parameters other than incidence to take on their actual values in each year,

[7] To fit the underlying time trend in incidence, the regressions also include two-part splines with a break point in 1992. Full details of these regressions are available in Liebman (2014).

while adjusting the incidence rate in each year using the coefficients from my regressions of incidence on the unemployment rate.[8] This modified beneficiary to working-age population series is used as the benchmark for the counterfactual simulations. In addition, when I conduct simulations holding incidence constant at the value from a base year, I hold it constant at the unemployment-adjusted value from that base year.

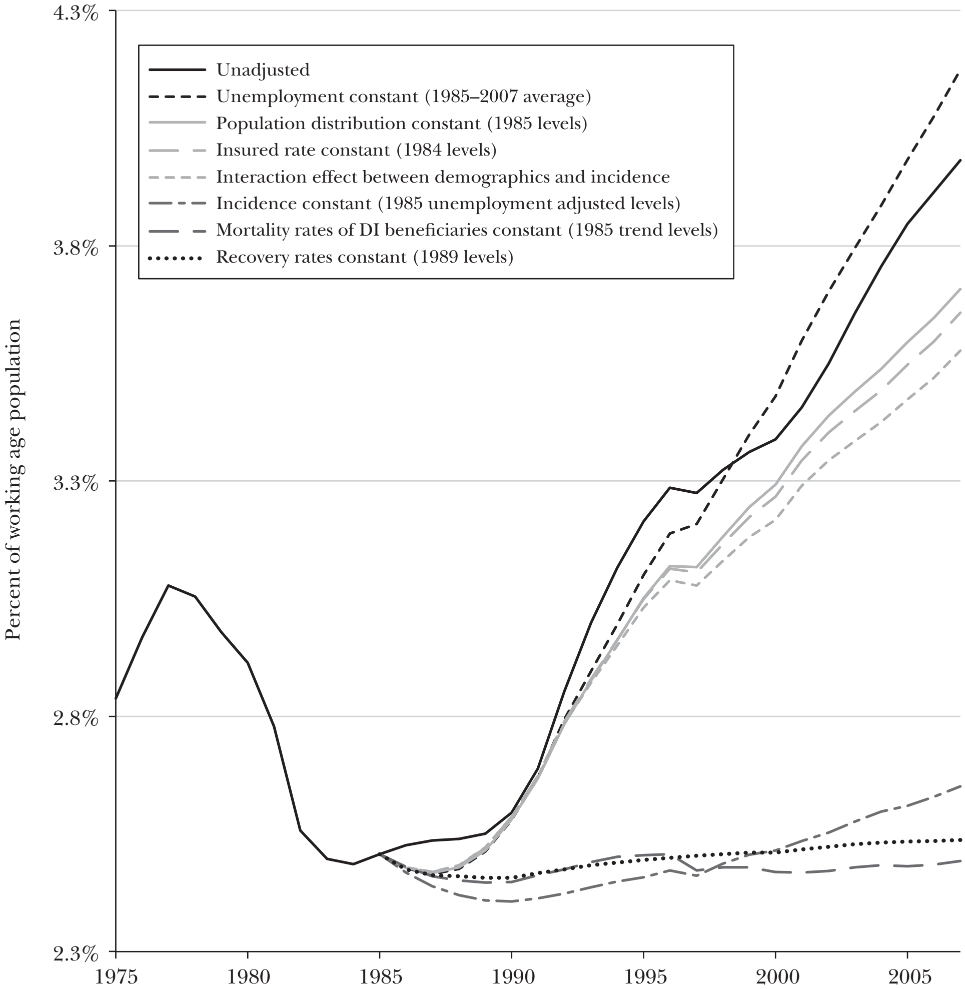
Men and women are analyzed separately, because of the very different evolution of their labor market experience in recent decades. Figure 5 shows the results of the simulations for men during the 1985–2007 period while Figure 6 will do the same for women. In Figure 5, the solid dark line shows the actual evolution of the men's DI beneficiary ratio, rising from 2.46 to 3.93 percent between 1985 and 2007. The rise in disability rates during the second half of the 1970s, the fall after the late 1970s and early 1980s policy changes, and the subsequent rise starting around 1985 all appear clearly. The next line in the key shows the beneficiary ratio with the actual incidence for each year adjusted to the value predicted if unemployment had remained steady at 5.6 percent in each year. Because the unemployment rate was relatively high for most of the late 1980s and early 1990s and low in the late 1990s and early 2000s, this unemployment-constant series is below the actual values in the early part of the analysis period and above it in the later period. The 2007 value for this adjusted series is 4.12 percent. The next line in the key holds the population age-distribution constant at its 1985 values (chosen because it is the initial year of the simulations). Absent the aging of the baby boomers, the DI beneficiary ratio in 2007 would have been 3.66 percent. The next line in the key shows that additionally holding the male "insured rate" constant at its 1984 level (chosen because it is approximately the average level in the 1985–2007 period) has little impact on the DI beneficiary level, reducing it only to 3.61 percent—because the share of males eligible for DI did not change much during most of this time period. To examine the impact of incidence, I adjust 1985 incidence to the value that my regressions predict would have occurred if unemployment had been 5.6 percent in that year; then I hold incidence constant at this unemployment-adjusted 1985 value (this is in addition to holding the age-distribution and insured rate constant). Doing so reduces the beneficiary to worker ratio in 2007 to 2.60. Compared to the insured-rate constant line, the reduction from 3.61 to 3.53 percent is attributable to the interaction effect between the demographic parameters and incidence, while the reduction from 3.53 to 2.60 percent is the direct effect of rising incidence if the population distribution and insured rate had not changed.

Holding mortality rates of DI beneficiaries constant—in addition to holding the earlier factors constant—further reduces the simulated 2007 Disability

[8] Specifically, I replace the incidence rate, I_{at}, for age a and year t, with an unemployment-adjusted incidence rate, $I_{at}^{*} = I_{at} + \beta_{gc}(\overline{U} - U_t) + \beta_{gl}\ (\overline{U} - U_{t-1})$, where β_{gc} and β_{gl} are the coefficients from the regression of incidence on contemporaneous and lagged unemployment for the 5-year age group that a belongs to. This assumes a simple additive relationship between changes in unemployment and incidence. It would be valuable to do additional research, perhaps using state-level data, into the best functional form for the relationship between unemployment and DI incidence.

Figure 5

Impact of Various Factors on the Percentage of Working Age Men (Ages 20–64) Receiving Disability Insurance, 1985–2007

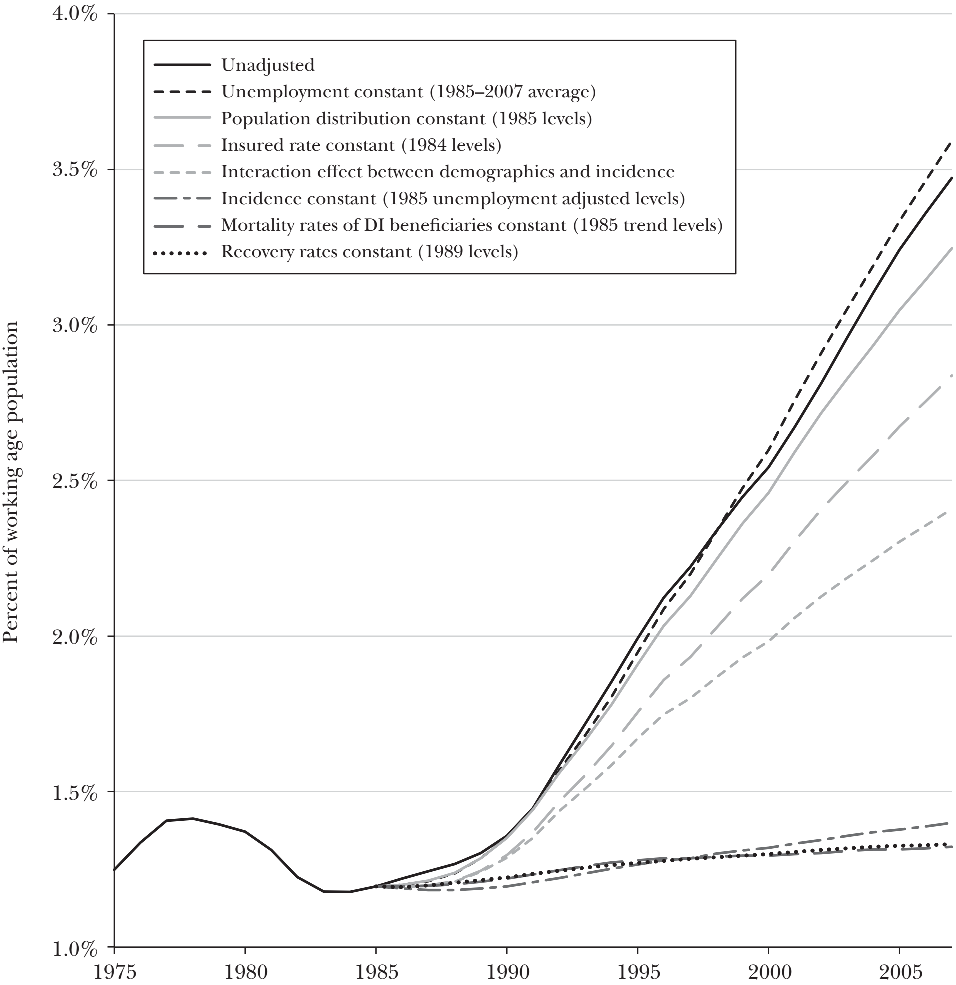


Sources: Social Security Administration, Office of the Chief Actuary; and author's calculations.

Notes: In this analysis, each factor is analyzed sequentially relative to all of the other factors that are listed before it in the key. Thus the vertical distance between a line and the line that comes before it in the key represents the additional effect of holding the factor constant on top of holding all of the earlier factors constant. I attribute to incidence the increase in enrollment rates that would have occurred absent population aging and changing insured rates. Separately, I estimate the effect of population aging and changing insured rates under a counterfactual scenario in which incidence rates remained constant. The sum of these separate estimates is smaller than the total effect when all three factors are held constant together. I classify the difference between the separate effects and the total effect as "interaction effects." I stack the two quantitatively less-important factors—mortality rates and recovery rates—at the end of the analysis and do not estimate separate interaction effects for them. Also, I first modify the actual beneficiary to working-age population ratio to provide an alternative series that projects the path that the ratio would have taken if the unemployment rate had remained steady at its average value for the entire time period analyzed for the simulation. The unemployment adjustment uses the mean unemployment and lagged (1 year) unemployment from 1985 to 2007. See text for details.

Figure 6

Impact of Various Factors on the Percentage of Working-Age Women (Ages 20–64) Receiving Disability Insurance, 1985–2007



Sources: Social Security Administration, Office of the Chief Actuary; and author's calculations.
Notes: In this analysis, each factor is analyzed sequentially relative to all of the other factors that are listed before it in the key. Thus the vertical distance between a line and the line that comes before it in the key represents the additional effect of holding the factor constant on top of holding all of the earlier factors constant. I attribute to incidence the increase in enrollment rates that would have occurred absent population aging and changing insured rates. Separately, I estimate the effect of population aging and changing insured rates under a counterfactual scenario in which incidence rates remained constant. The sum of these separate estimates is smaller than the total effect when all three factors are held constant together. I classify the difference between the separate effects and the total effect as "interaction effects." I stack the two quantitatively less-important factors—mortality rates and recovery rates—at the end of the analysis and do not estimate separate interaction effects for them. Also, I first modify the actual beneficiary to working-age population ratio to provide an alternative series that projects the path that the ratio would have taken if the unemployment rate had remained steady at its average value for the entire time period analyzed for the simulation. The unemployment adjustment uses the mean unemployment and lagged (1 year) unemployment from 1985 to 2007. See text for details.

Insurance beneficiary rate to 2.44 percent. Age-adjusted mortality rates for male DI beneficiaries fell from 4.9 percent in 1982 to 3.2 percent in 2007 a phenomenon that is discussed further below. In holding mortality rates constant, I use a base that is a weighted average of 1982 mortality rates and 1998 mortality rates, with 80 percent of the weight on the 1982 rates. Doing so provides a base level for 1985 that is on the longer-term mortality trend line, avoiding the spike in actual mortality that occurred after the removal of less-impaired individuals from the DI beneficiary rolls in the early 1980s and the spike in HIV-related mortality that begins in the 1980s and continues into the mid 1990s. In the final step, additionally holding "recovery rates"—that is, the rate at which eligibility for benefits terminates for a reason other than death, typically an improvement in health—at their 1989 level has only a small further impact on the simulated 2007 DI beneficiary rate, increasing it to 2.49 percent. 1989 was chosen because recovery rates were quite stable over the time period and it is the year with approximately the average recovery rate for the 1985–2007 period, excluding the one-year spike that occurred in 1997 when beneficiaries whose main impairment was related to drug or alcohol use were removed from the rolls.[9]

The left-most bar in Figure 7 and the first column of the top panel of Table 1 summarize the simulation results for men by showing the percentage of the distance from the 2007 unemployment-adjusted beneficiary ratio of 4.12 percent, to the simulated ratio of 2.49 percent with all of the factors held constant, that is attributable to each factor. For men over the 1985–2007 period, population aging is responsible for 28 percent of the increase in the DI beneficiary ratio. The insured rate is responsible for a negligible 3 percent. Actual incidence being above the 1985 unemployment-adjusted level is responsible for 57 percent, with the interaction between the demographic factors and incidence responsible for 5 percent. Falling death rates are responsible for 10 percent. The recovery rate being higher than the base value is responsible for −3 percent.

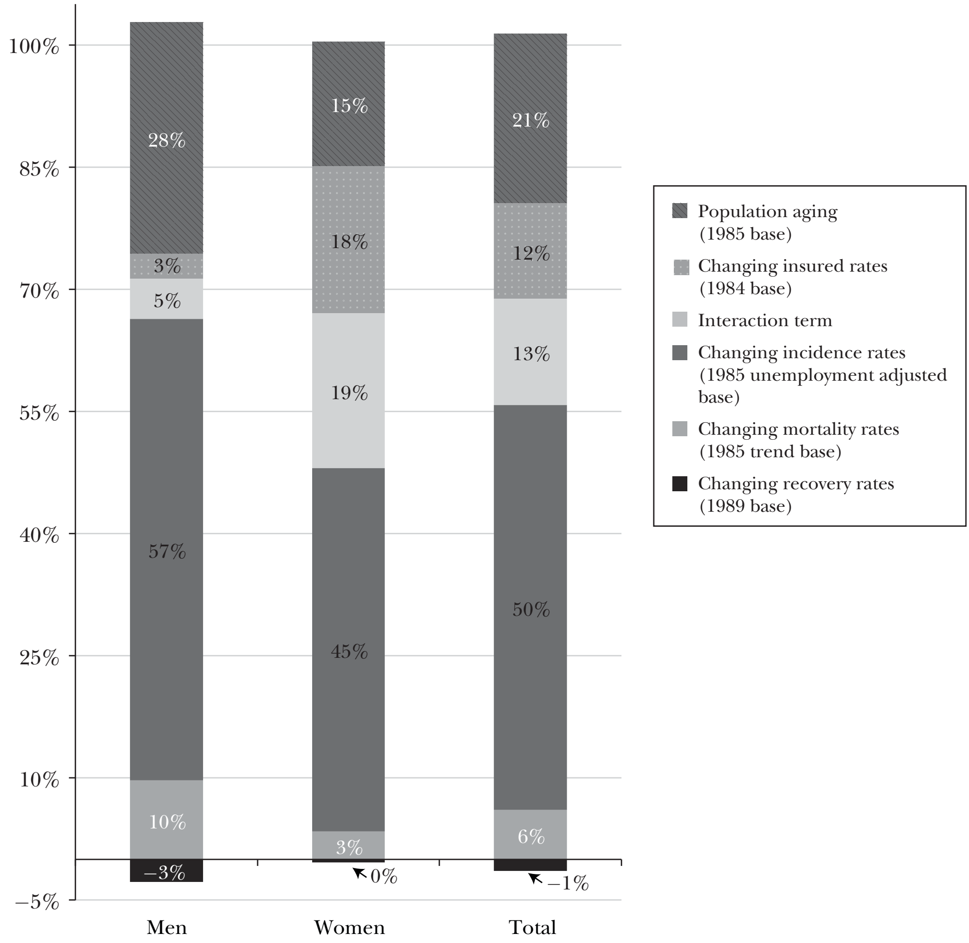
As I emphasized above, the decomposition results are highly sensitive to the incidence base year. Column 6 of the top panel of Table 1 shows that if I had begun the analysis in the high incidence year of 1977 (rather than the low incidence year of 1985) and studied the entire 1977–2007 period for men and women combined, I would have found that changing incidence reduced the DI enrollment rate and that population aging and rising insured rates each accounted for approximately half of the rise in enrollment over the 30-year period.

There have really been three distinct subperiods, as shown in the bottom panel of Table 1. From 1977 to 1985 the male beneficiary ratio fell sharply with falling incidence rates explaining 61 percent of the decline and higher recovery rates explaining 32 percent. From 1985–1993, rising incidence is responsible for 105 percent of the increase in male benefit receipt, while population aging is responsible for only 5 percent. Mortality rates exceeded their trend level during this

[9] Liebman (2014) contains additional details on the time-path of each of these parameters.

Figure 7

Decomposition of Various Factors' Impact on the Percent of the Working-Age Population Receiving Disability Insurance, 1985–2007



Sources: Social Security Administration, Office of the Chief Actuary; and author's calculations.

period, reducing benefit receipt and accounting for −13 percent. From 1993–2007, population aging accounts for 93 percent of the increase in benefit receipt for men and falling mortality rates account for 38 percent. Incidence was on average below its 1993 level and accounted for −25 percent of the increase for men. The spike in recovery rates from eliminating eligibility for those with impairments related to drug and alcohol addiction also contributed −25 percent. Given the result presented in Figure 4 that age- and unemployment-adjusted male incidence rates fell sharply in the early 1980s, rose steeply in the second half of the 1980s, and have been steady

Table 1
Decomposition of Various Factors' Impact on the Percent of the Working Age Population Receiving Disability Insurance

	A. Full time periods with two different base years					
	1985–2007			*1977–2007*		
	Men	*Women*	***Total***	*Men*	*Women*	***Total***
	(1)	*(2)*	***(3)***	*(4)*	*(5)*	***(6)***
Change in beneficiary ratio to be explained	1.6	1.6	**1.6**	0.4	1.8	**1.1**
Percent explained by:						
Population aging	28%	15%	**21%**	142%	29%	**52%**
Changing insured rates	3%	18%	**12%**	11%	58%	**48%**
Interaction term	5%	19%	**13%**	−19%	2%	**−2%**
Changing incidence rates	57%	45%	**50%**	−72%	6%	**−9%**
Changing mortality rates	10%	3%	**6%**	53%	5%	**15%**
Changing recovery rates	−3%	0%	**−1%**	−15%	−1%	**−4%**

	B. Three subperiods								
	1977–1985			*1985–1993*			*1993–2007*		
	Men	*Women*	***Total***	*Men*	*Women*	***Total***	*Men*	*Women*	***Total***
	(1)	*(2)*	***(3)***	*(4)*	*(5)*	***(6)***	*(7)*	*(8)*	***(9)***
Change in beneficiary ratio to be explained	−1.4	−0.6	**−1.0**	0.4	0.4	**0.4**	0.3	1.0	**0.7**
Percent explained by:									
Population aging	4%	4%	**4%**	5%	4%	**4%**	93%	27%	**44%**
Changing insured rates	1%	−20%	**−5%**	−1%	26%	**13%**	12%	24%	**21%**
Interaction term	−1%	7%	**1%**	2%	10%	**6%**	5%	6%	**6%**
Changing incidence rates	61%	71%	**64%**	105%	64%	**84%**	−25%	39%	**23%**
Changing mortality rates	4%	5%	**4%**	−13%	−5%	**−9%**	38%	7%	**15%**
Changing recovery rates	32%	32%	**32%**	2%	1%	**1%**	−25%	−3%	**−9%**

Sources: Social Security Administration, Office of the Chief Actuary; and author's calculations.
Notes: I first modify the actual beneficiary-to-working-age population ratio to provide an alternative series that projects the path that the ratio would have taken if the unemployment rate had remained steady at its 1985–2007 mean value of 5.6 percent. I then hold factors constant, using the same sequential method as in Figures 5 and 6. Each column represents one run of the model, where the top row gives the difference in percentage points (for the final year of the simulation time period) between the alternative beneficiary ratio, which holds only the unemployment rate constant, and the last counterfactual beneficiary ratio, which holds all factors constant. The other rows represent the percent of this difference that can be attributed to each factor, including the interaction between incidence and population factors (aging and insured rates). The effect of population aging is found by holding the population age distribution constant at its distribution in the starting year for each model run (1977, 1985, or 1993). Similarly, I hold insured rates constant at their values in each of the three start years and hold incidence rates constant at their unemployment-adjusted values in each of the three start years. For mortality rates, I attempt to find values on the long-term trend line, so that my results are not distorted by the spike in actual mortality that occurred after the removal of less-impaired individuals from the DI beneficiary roles in the early 1980s and by the high rate of mortality among men with HIV in the 1980s and early 1990s. Therefore, I hold mortality constant at 1977 values in the model runs that begin in 1977; at a 1985-trend value, which reflects a weighted average of 1982 and 1998 mortality rates, in the model runs that begin in 1985; and at a 1993-trend value, which is found by averaging 1996 and 1997 mortality rates, in the model runs that begin in 1993. Recovery rates are the final factor I hold constant, and I do so at 1989 values for all three scenarios, because recovery rates have remained quite stable over time, and 1989 is the year with approximately the average recovery rate for the 1985–2007 period, excluding the one-year spike that occurred in 1997 when beneficiaries whose main impairment was related to drug or alcohol use were removed from the rolls.

(or falling slightly) since the early 1990s, this time pattern of results should not be surprising.[10]

I next perform an analogous set of counterfactual simulations that ask how the Disability Insurance beneficiary ratio would have evolved for women, holding constant the various factors for the 1985–2007 period. The base years used for each factor are the same as they were for men.

In Figure 6, the dark line shows the actual evolution of the female Disability Insurance beneficiary ratio, rising from 1.20 to 3.47 percent between 1985 and 2007. The next line in the key shows the adjusted ratio, with unemployment held constant. As with men, this results in a series that is somewhat lower in the first half of the period and somewhat higher in the second half. The 2007 value of this series is 3.59 percent. The next line in the key, additionally, holds the population age-distribution constant at its 1985 level. Absent the aging of the baby boomers, the female DI beneficiary ratio in 2007 would have been 3.25 percent. The next line in the key shows that additionally holding the female insured rate constant at its 1984 level has a fairly large impact on the beneficiary rate—lowering it to 2.84 percent. This factor is larger for women than for men because of the large-scale entry of women into the workforce starting in the 1970s that has resulted, over time, in a much larger share of women being covered by disability insurance. Additionally, holding incidence at its 1985 unemployment-adjusted average reduces the simulated beneficiary rate to 1.40 percent, with 30 percent of the reduction resulting from the interaction effect. Holding mortality rates constant at their 1982 level, on top of holding all of the earlier factors constant, has a somewhat smaller impact than for men because female mortality is lower; it reduces the simulated 2007 Disability Insurance beneficiary rate to 1.32 percent. Additionally holding recovery rates at their 1989 level has essentially no further impact on the simulated 2007 DI beneficiary rate.

Figure 7 and column 2 of the top panel of Table 1 summarize these results, showing that for women population aging and rising insured rates combine to account for one-third of the increase in the beneficiary ratio over the entire 1985 to 2007 period. Rising incidence accounts for 45 percent, and the interaction between the demographic factors and rising incidence accounts for 19 percent. The impact of changes in mortality and recovery rates was negligible.

The decomposition of results by subperiod in Table 1 shows that the time pattern of results for women is somewhat different from that of men, primarily because rising insured rates are a more significant factor for women. From 1977–1985, falling incidence rates explain 71 percent of the decline in enrollment

[10] The change in the beneficiary ratio for the 1985–2007 and 1977–2007 periods is greater than the sum of the changes in that ratio for the relevant subperiods. This occurs because it can take decades to reach a new steady state beneficiary ratio after, for example, a change in the incidence rate. Thus the increase in incidence after 1985 was still causing the beneficiary ratio to rise throughout the 1990s when compared to a 1985 incidence base, and this is reflected in the simulations for the full 1985–2007 period. But the impact of the 1980s increase in incidence is not captured in the simulations for the 1993–2007 subperiod, which use a 1993 incidence base and reflect only the impact of further changes in incidence relative to the 1993 level.

for women, rising recovery rates explain 32 percent, and rising insured rates (which increase enrollment) are responsible for -20 percent. Whereas rising incidence accounted for nearly all of the increase in the beneficiary ratio for men in the 1985–1993 time period, for women 64 percent of the increase was the result of rising incidence and 26 percent was the increase in insured rates. Whereas population aging and declining mortality rates accounted for nearly all of the increase in the male beneficiary ratio for the 1993–2007 time period, for women population aging, rising insured rates, and rising incidence all played a role. In particular, as we saw in Figure 4, incidence rates for women have continued to rise even after those for men leveled off.[11]

The impression in policy circles that disability enrollment and spending are "out of control" appears to be the result of confounding the legislatively induced bounce-back of incidence rates in the late 1980s and early 1990s with the largely demographically induced increases of the past two decades. There have been three different phenomenon, each with its own time path and economic origins. The first is a legislatively induced rise in disability incidence rates that explains the bulk of program growth between 1985 and the early 1990s. The second is rising female labor force participation, which enabled a greater share of women to qualify for SSDI benefits. The third factor, and the largest contributor to rising SSDI rolls between the early 1990s and the onset of the Great Recession, is the entry of the baby boom generation into its peak disability years. All three factors have now arguably run their course in terms of increasing the share of GDP spent on DI benefits. But changes in the characteristics of the beneficiary population in recent decades could augur future changes in the program. I turn to this subject next.

Changes in the Beneficiary Population

Much of the policy attention to the Disability Insurance program is motivated by a concern that higher enrollment rates may be the result of an expansion in benefit receipt by individuals with less-severe impairments. According to this perspective, the 1984 legislative reforms and the way in which they have been administered loosened eligibility criteria, and the impact of the altered eligibility standards was magnified by challenging labor market conditions for low-skilled workers, which increased their incentive to claim benefits.

While it is difficult to directly observe whether eligibility standards have shifted over time, we can find clues by looking at trends in the age distribution of claims, the medical impairments triggering eligibility, and the mortality rate of beneficiaries. Such clues need to be interpreted with care. One cannot assess the standards

[11] These results attribute a larger share of the increase in DI enrollment to demographic factors than do Duggan and Imberman (2009), who examine the period 1984–2003. They attribute 15 percent of the rise in enrollment among men and 4 percent of the rise among women to changes in the age structure of the population.

applied to disability benefits simply by looking at the age-adjusted rates of disability incidence, because incidence rates are affected by factors beyond how the program is administered. For example, declining relative demand for low-wage workers and stagnating real wages at the bottom of the income distribution increased the incentives for low-skill workers to apply for disability benefits during the 1980s and 1990s (Autor and Duggan 2003). These changes in incentives would be predicted to increase the rate of disability benefit claiming, which suggests that stable disability incidence rates in the post-1990 period could be indicative of tighter eligibility standards being applied. Conversely, if the overall health of the population is improving, then we would expect declining incidence of disability, and a finding of stable incidence rates could reflect looser eligibility standards. Moreover, greater take-up of disability insurance in an era of declining economic prospects for low-skilled workers could be socially optimal since the economic cost of workers foregoing labor force participation depends on the marginal product of their labor relative to their disutility of work (Diamond and Sheshinski 1995).

Some observers have cited a shift in the age composition of the disability beneficiary population toward younger ages as evidence that disability determination standards have become more lenient. Among both men and women, the mean age of new beneficiaries fell by more than three years between 1980 and 1993. However, between 1993 and 2011, the mean age of new beneficiaries increased by three years, returning to early 1980s levels. The complication in interpreting these trends is that as the baby boomers moved through their life cycle, they first swelled the number of younger workers, which mechanically increased the share of younger workers claiming disability benefits, and then later increased the share of older disability claimants. Indeed, when the ages of new recipients of disability benefits are adjusted to hold the age composition of the insured population constant, the average age fell significantly from the early 1980s to the early 1990s, but has fluctuated around a relatively stable trend since 1990. This pattern is consistent with an interpretation that eligibility standards expanded significantly in the aftermath of the 1984 legislation, but have been relatively stable since the early 1990s.

Another piece of evidence comes from examining the incidence of specific medical impairments. The stability of the overall (age- and unemployment-adjusted) disability incidence rate in the post-1990 period masks substantial changes in the incidence of individual impairments. For both males and females, the incidence of circulatory- and cancer-related benefit awards has been falling, while the incidence of musculoskeletal and, to a lesser extent, mental conditions has been rising. One possible interpretation of these trends is that overall health has been improving as reflected in the declining circulatory and cancer incidence rate, but that improving health has not produced declining incidence rates because the program has become more lenient in approving claims for musculoskeletal and mental conditions. Using my simulation model, I find that if the incidence rates for musculoskeletal and mental benefit awards had remained constant at their 1985 levels, while all other conditions followed their actual path, the beneficiary ratio would have been 21 percent lower in 2007 than it was.

However, there are other possible interpretations for the shift in the distribution of impairments. For example, it could be that standards for determining disabilities have remained constant, but that a greater number of individuals with musculoskeletal or mental health conditions have applied for benefits, either because the prevalence of the conditions has increased over time or, more likely, because labor market conditions for low-skilled workers have increased the incentives for individuals with these conditions to apply for benefits. It is also possible that some of the shift in the distribution of impairments was the result of individuals who would have been eligible for benefits under other categories (possibly a few years later) instead claiming benefits under the musculoskeletal and mental impairment categories after the 1984 reforms made such claims easier.

The fact that the relatively stable rates of (adjusted) disability incidence during the past 25 years were the result of large offsetting trends in incidence rates for different conditions suggests that there should be no presumption that rates will be stable going forward. For example, if incidence rates for musculoskeletal and mental health impairments continue to rise, but the offsetting declines in the other conditions level off, overall incidence could rise. Relatedly, while female disability incidence rates have leveled off since the mid-1990s at a rate slightly below male rates, giving the appearance that the earlier rapid rise in female incidence rates was largely a phenomenon of female rates converging to male rates as female labor market behavior became more similar to male behavior, incidence rates for particular conditions are quite different for men and women, suggesting that the appearance of convergence in the aggregate patterns may simply be a coincidence.

A final piece of evidence comes from mortality rates among Disability Insurance recipients. These rates have continued to fall, even during the period in which adjusted incidence rates have mostly stabilized. This observation is consistent with an interpretation that there has been a shift in the composition of disability beneficiaries toward impairments like musculoskeletal and mental impairments that have lower mortality rates. Although it is conceivable that medical progress has significantly reduced mortality for a wide range of conditions without improving functional capacity, it seems likely that a significant portion of the decline in mortality rates among DI recipients is the result of a change in the composition of the beneficiary population.

Priorities for Reform of Disability Insurance

By international standards, US spending on disability benefits relative to GDP remains low. The OECD provides data on total public expenditures on disability and sickness cash benefits for its member countries. In 2011, average spending in the OECD on these benefits was 1.9 percent of GDP. In the US, it was 1.3 percent of GDP. The Netherlands, a country often heralded for its aggressive disability benefit reforms, spent 2.8 percent of GDP on these benefits in 2011 (down from 6.5 percent in 1980). Despite the relatively modest US expenditures on these programs, there is a strong case for treating the coming exhaustion of the Disability Insurance trust fund as an opportunity for improving the US Disability Insurance system.

Social insurance programs need to be designed to balance the protection they provide with the economic distortions they cause (Feldstein 1976). Disability insurance benefits provide protection against the risk of a severe medical impairment, while they also generate disincentives for labor force participation. But economic research suggests that some significant aspects of the disability insurance system are so far from the optimal policy frontier that reforms may exist that can simultaneously improve the well-being of impaired individuals and reduce the fiscal and efficiency costs of the program.

Improved Incentives for Returning to Work

The current disability benefit package essentially provides lifetime cash benefits and health insurance in exchange for a promise never to do substantial work again. That is, given that only about 1 percent of beneficiaries per year are removed from the rolls based on health improvements, so long as a beneficiary does not have significant labor earnings, the individual is unlikely to lose eligibility for benefits. A sizable portion of the disabled beneficiary population might be better off with assistance that helps them return to employment. Changes in the disability insurance programs and in low-skill labor markets, along with the decline in other forms of public assistance, have made this group a larger fraction of the Disability Insurance and Supplemental Security Income population (Autor and Duggan 2003).

The evidence that a significant number of disability beneficiaries have the capacity to work comes from a line of research that began with Bound (1989) and examines the earnings of applicants who are denied disability benefits to assess the earnings potential of marginal beneficiaries.[12] A welcome evolution in this literature uses the random assignment of disability cases to examiners or Administrative Law Judges with different propensities to approve awards to generate a causal estimate of the effect of Disability Insurance awards on labor supply (Autor, Maestas, Mullen, and Strand 2015; French and Song 2011). It also accounts for the fact that the lengthy DI application process can erode labor force participation even among applicants who are eventually denied disability benefits (Maestas, Mullen, and Strand 2015). This literature finds that applying for and receiving DI reduces employment rates by over 30 percentage points overall and by more than 50 percentage points among those with lesser impairments. Roughly one-quarter of applicants are on the margin of program entry in the sense that they receive benefits if their case is assigned to a lenient examiner, but not if they are assigned to one with a lesser propensity to award benefits (Maestas, Mullen, and Strand, 2013). However, the subsequent earnings levels of denied applicants who return to employment are generally below $20,000, suggesting that without further assistance

[12] See von Wachter, Song, and Manchester (2011) for a more recent application of the Bound (1989) methodology and Moore (2015) for an analysis of the impact of terminating DI benefits on subsequent labor supply.

the labor market prospects of individuals on the margin between receiving and not receiving benefits is quite limited.

Incentives for Employers, States, and the Social Security Administration

Several of the key actors in the disability insurance system have misaligned incentives that cause them to encourage people to apply for disability insurance (Liebman and Smalligan 2013). A number of the ideas for reform of the US Disability Insurance system seek to alter these incentives.

For example, when an employee experiences a health problem, an employer may find it easier and less expensive to push an employee toward applying for Disability Insurance benefits than to make accommodations that would allow the worker to remain employed at the firm. Similarly, it is often less expensive for private disability insurance companies to help workers sign up for public Disability Insurance benefits than to help them get back to work.

Several reform proposals target incentives for employers, in part based on the observation that intervening early, before someone becomes detached from employment, is more effective than trying to connect someone later to a new job. For example, Autor and Duggan (2010) propose that employers be required to provide private disability insurance coverage to all of their workers and that this insurance would cover the first two years of a person's disability. Eligibility for federal benefits would begin only after the two years of private benefits were exhausted. In their formulation, benefits would be 60 percent of prior earnings and would also include vocational rehabilitation and workplace accommodations. Because employers would be charged different rates by the private insurance companies depending on the benefit claims of their employees, employers would have an incentive to find ways to keep their disabled workers employed. In order to create greater incentives for firms to retain workers with health impairments, Burkhauser and Daly (2011) propose experience rating for the employer share of Disability Insurance taxes in a way that is analogous to how worker's compensation and unemployment insurance contributions are experience rated. Thus, if an employer had a larger number of its workers claiming disability, that employer would face higher Disability Insurance premiums.

Other important decision makers who affect whether workers end up receiving Disability Insurance, or not, include states and the Social Security Administration itself. States have incentives to encourage low-wage workers to sign up for Disability Insurance and Supplemental Security Income because doing so has the effect of shifting both cash assistance costs and health care costs to the federal government and away from state programs. A change in federal funding formulas could alter this incentive.

The Social Security Administration's administrative budget comes from capped discretionary spending while benefits are mandatory. As a result, the Social Security Administration often ends up underinvesting in administrative capacity—failing to do continuing disability reviews, for example—even when doing so increases total program costs. Thus, the Social Security Administration has a backlog of 1.4 million continuing disability reviews even though its actuaries estimate that every $1 spent on continuing disability reviews saves $10 in future benefits (Social Security Administration 2013). Additional administrative capacity would lead to more timely and accurate initial disability decisions, possibly reducing the number of cases that are appealed. In Liebman and Smalligan (2013), we propose that the funding for state disability determination services be switched to the mandatory side of the budget, which would be in accord with how the administrative costs of TANF, Medicaid, and Food Stamps operate.

A Pilot Program Approach

In most cases, we lack the evidentiary base necessary to judge whether specific disability insurance reforms would do more good than harm. Are the earnings gains that can be produced from employment supports for partially disabled workers sufficient to be cost effective when compared with simply providing cash transfers? Would experience-rating of Disability Insurance benefits discourage firms from hiring either disabled workers or workers from demographic groups with higher incidence of disability? In Liebman and Smalligan (2013), we propose three federal pilot demonstrations to generate the needed learning. Because research has consistently shown that it is far less effective to intervene after a person has begun receiving disability insurance benefits, all of the pilots would be early intervention programs.

A first pilot program would test whether employer incentives can reduce Disability Insurance enrollment. Specifically, we propose a demonstration program that would provide a tax credit against firm DI payroll tax for firms that can reduce the disability incidence of their employees by at least 20 percent. A second demonstration would screen disability applicants and target those who appear likely to be determined eligible for benefits but who also have the potential for significant work activity if provided with a proper range of services. In exchange for suspending their disability insurance application, these applicants would be offered a package of benefits including targeted vocational and health interventions, a wage subsidy, and perhaps a few months of an emergency cash diversion grant. In this way, the demonstration would find out whether it is possible to improve the well-being of applicants while simultaneously achieving near-term cost neutrality and long-term savings. The third demonstration would allow several states to reorganize existing funding streams to target populations that are likely to end up receiving a lifetime of DI or Supplemental Security Income benefits in the absence of assistance. States would receive incentive funding if they demonstrate success at improving outcomes and reducing participation in DI and SSI. Similarly, Mann and Stapleton (2011) propose state-based disability insurance pilots analogous to the welfare waiver experiments of the 1980s and 1990s that informed the 1996 federal welfare reform.

As the Disability Insurance Trust Fund heads toward exhaustion in 2016, legislative action of some sort will be necessary. While it is possible to delay substantive changes to the DI program for another decade or more simply by raising the share of the OASDI payroll tax that is directed to the DI trust fund and lowering the share that is directed to the retirement trust fund, more significant changes will ultimately be needed. It would be wise, therefore, for the upcoming legislation to authorize a series of demonstration projects that can increase the chance that when it becomes time for more significant reforms, we will know enough to make smart choices. Economic research over the past two decades has suggested a set of changes that, by addressing some of the misplaced incentives in the system, offer the possibility of saving funds in the disability insurance system while potentially making people better off. These changes include altering the disability benefit package in a way that focuses on helping a larger proportion of the disabled return to work and reforming misaligned incentives that currently lead firms and state governments to encourage too many people to apply for federally funded disability benefits. It will take additional creative economic thinking in the next few years to design and evaluate the research and pilot projects that are needed to provide the evidence to guide broader reforms.

■ *The author thanks Wayne Sandholtz and Emily Tisdale for excellent research assistance.*

References

Autor, David H., and Mark G. Duggan. 2003. "The Rise in the Disability Rolls and the Decline in Unemployment." *Quarterly Journal of Economics* 118(1): 157–206.

Autor, David H., and Mark G. Duggan. 2006. "The Growth in the Social Security Disability Rolls: A Fiscal Crisis Unfolding." *Journal of Economics Perspectives* 20(3): 71–96.

Autor, David H., and Mark G. Duggan. 2010. "Supporting Work: A Proposal for Modernizing the US Disability Insurance System." Center for American Progress and The Hamilton Project.

Autor, David H., Nicole Maestras, Kathleen J. Mullen, and Alexander Strand. 2015. "Does Decay Cause Decay? The Effect of Administrative Decision Time on the Labor Force Participation and Earnings of Disability Applicants." NBER Working Paper 20840.

Bound, John. 1989. "The Health and Earnings of Rejected Disability Insurance Applicants." *American Economic Review* 79(3): 482–503.

Burkhauser, Richard V., and Mary Daly. 2011. *The Declining Work and Welfare of People with Disabilities.* AEI Press.

Congressional Budget Office (CBO). 2012. "Policy Options for the Social Security Disability Insurance Program." Report, July 16.

Cutler, David, Angus Deaton, and Adriana Lleras-Muney. 2006. "The Determinants of Mortality." *Journal of Economic Perspectives* 20(3): 97–120.

Cutler, David M., Ellen Meara, and Seth Richards-Shubik. 2012. "Unemployment and Disability: Evidence from the Great Recession." NBER Retirement Research Center Paper NB 12-12.

Diamond, Peter, and Eytan Sheshinski. 1995. "Economic Aspects of Optimal Disability Benefits." *Journal of Public Economics* 57(1): 1–23

Duggan, Mark, and Scott A. Imberman. 2009. "Why Are the Disability Rolls Skyrocketing? The Contribution of Population Characteristics, Economic Conditions, and Program Generosity." Chap. 11 in *Health at Older Ages: The Causes and*

Consequences of Declining Disability among the Elderly, edited by David Cutler and David Wise. University of Chicago Press.

Feldstein, Martin. 1976. "Seven Principles of Social Insurance." *Challenge* 19(5): 9–11.

Feldstein, Martin S., and Jeffrey B. Liebman. 2002. "Social Security." *Handbook of Public Economics,* vol. 4, pp. 2245–324. Amsterdam: Elsevier.

French, Eric, and Jae Song. 2014. "The Effect of Disability Insurance Receipt on Labor Supply." *American Economic Journal: Economic Policy* 6(2): 291–337.

Kearney, John R. 2005/2006. "Social Security and the 'D' in OASDI: The History of a Federal Program Insuring Earners against Disability." *Social Security Bulletin* vol. 66, no. 3.

Liebman, Jeffrey B. 2014. "Understanding the Increase in Disability Insurance Spending." NBER Disability Research Center Working Paper NB 13-01.

Liebman, Jeffrey B., and Jack A. Smalligan. 2013. "An Evidence-Based Path to Disability Insurance Reform." The Hamilton Project, Brookings Institution.

Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand. 2013. "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *American Economic Review* 103(5): 1797–1829.

Mann, David R., and David C. Stapleton. 2011. "Fiscal Austerity and the Transition to Twenty-First Century Disability Policy: A Road Map." Mathematica Policy Center for Studying Disability Policy.

Mashaw, Jerry L. 1983. *Bureaucratic Justice: Managing Social Security Disability Claims.* Yale University Press.

Meuller, Andreas I., Jesse Rothstein, Till M. von Wachter. 2013. "Unemployment and Disability Insurance in the Great Recession." NBER Working Paper 19672.

Moore, Timothy. 2015. "The Employment Effects of Terminating Disability Benefits." Melbourne Institute Working Paper No. 2/15.

Muller, L. Scott. 2008. "The Effects of Wage Indexing on Social Security Disability Benefits." *Social Security Bulletin* vol. 68, no. 3.

OASDI Board of Trustees. 2014. *The 2014 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds.*

Reno, Virginia P. 2011. "Securing the Future of the Social Security Disability Insurance Program." Testimony before the House Committee on Ways and Means, Subcommittee on Social Security. December 2. http://www.nasi.org/research/2011/testimony-virginia-p-reno-securing-future-social-security-di.

Rupp, Kalman. 2012. "Factors Affecting Initial Disability Allowance Rates for the Disability Insurance and Supplemental Security Income Programs: The Role of the Demographic and Diagnostic Composition of Applicants and Local Labor Market Conditions." *Social Security Bulletin* vol. 72, no. 4.

Rutledge, Matthew S. 2011. "The Impact of Unemployment Insurance Extensions on Disability Insurance Application and Allowance Rates." Center for Retirement Research at Boston College Working Paper 2011-17.

Social Security Administration. 1998–2013. *Annual Statistical Supplement to the Social Security Bulletin.*

Social Security Administration. 2014a. *Annual Statistical Report on the Social Security Disability Insurance Program, 2013.*

Social Security Administration. 2014b. "Social Security Basic Facts." Webpage, April 2, 2014, http://www.ssa.gov/news/press/basicfact.html.

Social Security Technical Advisory Panel. 2011. *2011 Technical Panel on Assumptions and Methods.* Report to the Social Security Advisory Board.

von Wachter, Till, Jae Song, and Joyce Manchester. 2011. "Trends in Employment and Earnings of Allowed and Rejected Applicants to the Social Security Disability Insurance Program." *American Economic Review* 101(7): 3308–29.

Journal of Economic Perspectives—Volume 29, Number 2—Spring 2015—Pages 151–172

The Rise and Fall of Disability Insurance Enrollment in the Netherlands[†]

Pierre Koning and Maarten Lindeboom

For most economists, "Dutch disease" refers to the problems that economies often face in their manufacturing or export sector when there is a sharp increase in the development of energy or other natural resources. The phrase originally referred to how the manufacturing sector of the Netherlands was adversely affected by discoveries of natural gas in the late 1950s and has become a catch-all term for the difficulties experienced by many economies with high levels of natural resource exports. But for many European labor economists, "Dutch disease" also has another meaning. It refers to the fact that the share of those in the Netherlands who received disability benefits tripled from 4 percent of those who were insured in the late 1960s to about 12 percent of those who were insured in the mid-1980s—and then remained more or less constant at this unprecedented level until the beginning of the 21st century. As recently as 15 years ago, this high level of Disability Insurance (DI) enrollment was considered to be one of the major social and economic problems of the Netherlands; indeed, the Netherlands was characterized as the country with the most out-of-control disability program of OECD countries (Burkhauser, Daly, and de Jong 2008).

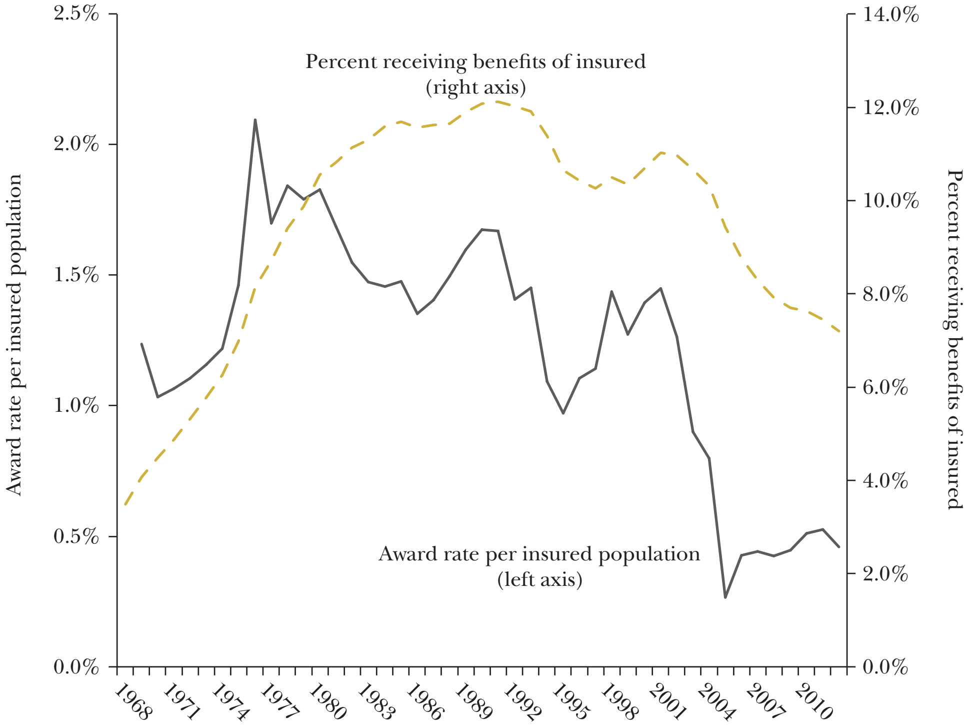
But since about 2002, the Netherlands has seen a spectacular decline in its Disability Insurance enrollment rate. Figure 1 shows the rise and fall. The share of the insured population receiving Disability Insurance decreased from 11 percent

■ *Pierre Koning is Part-time Professor of Labor Market and the Welfare State, and Maarten Lindeboom is Professor of Economics, both at VU University Amsterdam, Amsterdam, Netherlands. Both authors are also Research Fellows, Tinbergen Institute, Amsterdam, Netherlands, and Research Fellows, IZA (Institute for the Study of Labor), Bonn, Germany. Their email addresses are p.w.c.koning@vu.nl and m.lindeboom@vu.nl.*

[†] To access the Data Appendix, visit
http://dx.doi.org/10.1257/jep.29.2.151

doi=10.1257/jep.29.2.151

Figure 1
Disability Insurance Award and Enrollment Rates per Insured Worker in the Netherlands, 1968–2012



Source: UWV (2012).
Note: The Disability Insurance award rate is the share of the insured population that started to receive disability payments in a given year.

in 2001 to 7.2 percent in 2012. Similarly, the Disability Insurance award rates—that is, the share of the insured population that started to receive disability payments in a given year—declined from 1.5 percent in 2001 to about 0.5 percent in 2012. Also, spending on disability programs in the Netherlands halved from 4.2 percent of the GDP in 1990 to 2.1 percent in 2007 (OECD 2010). This rate of spending on disability benefits is lower than in comparable countries like Sweden (2.2 percent of GDP) and Norway (2.5 percent). In recent years, the number of disability beneficiaries per worker in the Netherlands has decreased below the level of the beneficiaries per worker for Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) in the United States.

A first question we address is what aspects of the program contributed to the increase of the disability rolls in the Netherlands until 2002. In brief, the disability program was set up in a way that caused it to function as an attractive substitute pathway into unemployment insurance for both workers and employers. Indeed, from the perspective of workers, DI benefit conditions remained generous until

2006, particularly compared to the benefits that could be received from unemployment insurance and social assistance benefit schemes. Similarly, during that time, employers had reason to prefer that some of their workers would be awarded disability benefits instead of unemployment benefits because in the Netherlands this avoided substantial firing costs. In the context of the Netherlands' broad disability scheme, which insures all workers against all income losses due to both occupational and nonoccupational injuries, workers and employers had ample opportunities to take advantage of the system. The disability scheme came to function like a long-term program for workers who were less employable rather than being restricted to those having substantial health problems.

Next we turn to reforms of the disability system in the Netherlands undertaken from 1996 to 2006. We cluster these reforms in three broad categories: 1) reducing the incentives of employers to move workers to disability; 2) increased gatekeeping; and 3) tightening disability eligibility criteria while enhancing worker incentives. As we will show, changes in the screening process and increased employer incentives have both contributed to a substantial decrease in inflows to disability benefits. However, changes in the duration and level of disability insurance benefits have had less effect. As it turns out, a key to the Dutch disability insurance reform has been transferring certain costs and responsibilities to employers, thus changing their incentives. In the Dutch system, workers are first placed on sick leave for two years before they become eligible for disability benefits. During that time, employers have become responsible for the continued payment of wages for two years of sickness, while disability benefit costs are—with some delay—passed on to employers by experience-rated premiums.

While these reforms are generally perceived as successful, there is also new criticism and concerns regarding some aspects of the current Disability Insurance program. The biggest concern is with the high level of sickness and DI risks that are transferred to employers, which probably has made employers more reluctant to hire workers with discernible health conditions. Although rigorous evidence in this direction is still limited, we will discuss whether employers have increased the screening and sorting of such workers. Related to this point, we should highlight the increased DI inflow rates of workers with flexible and/or temporary jobs. For these jobs, the DI benefit costs are not passed on to employers on an experience-rated basis.

Putting the Dutch Disability Insurance reforms in a broader perspective, a pertinent question is whether the dramatic decrease in the inflow rate to disability benefits was accompanied by increased employment rates, or whether those who would have been identified as disabled just ended up in other public support programs. To shed more light on these issues, we use survey data on the health status of individuals to investigate how differences in employment rates between people with good and bad health has evolved since the disability reforms took place. In light of the stricter eligibility criteria for disability that resulted from the reforms, it is likely that workers with bad health conditions are awarded DI benefits less frequently in the new scheme. According to the data, the reforms probably enhanced the work continuation of male workers with poor health to some extent. From this perspective, one may conclude that distortion in the labor supply of workers has decreased. At the same time, however, the share of unhealthy workers without work and receiving no disability benefits has increased. It thus is hard to infer whether the reforms in sum have contributed to the targeting efficiency of the DI program.

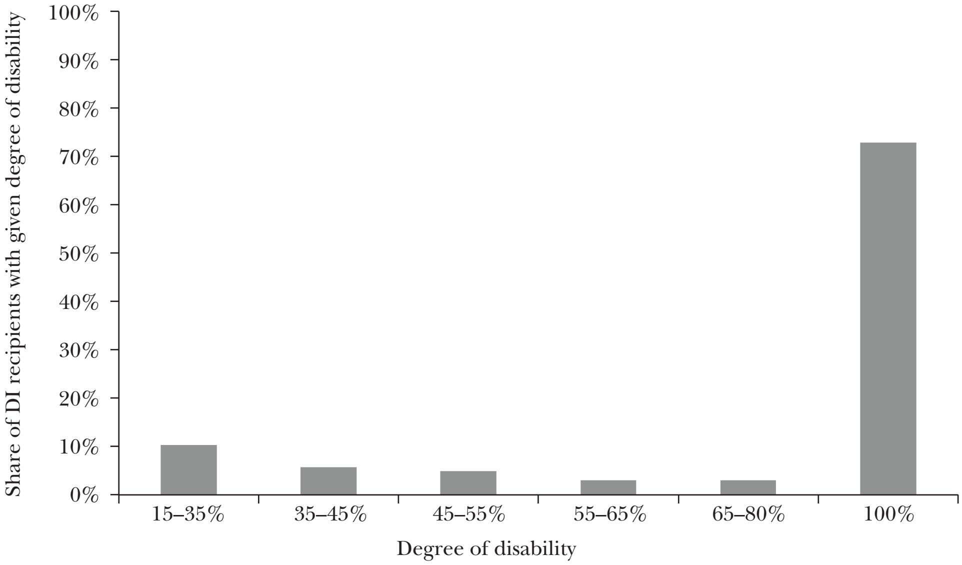
In the final section, we briefly summarize the main lessons that can be drawn from the reforms and discuss the major challenges the Dutch Disability Insurance system is facing in the years to come. Regarding the design of disability reforms in comparable industrialized countries, probably the most important lesson is that employers should be stimulated and facilitated in finding ways to prevent long-term sickness and absence, and subsequent disability inflow. The experiences with intensified gatekeeping during the sickness period show that employers can be pushed to take on this role. Indeed, the success of the Dutch disability reforms largely depends on the use of early interventions when a worker becomes sick, in the waiting period before they enter the disability rolls. At some point, however, employer obligations may become too sizeable, raising questions about the ability of employers to influence DI risks. Also if the obligations are too large, there is the risk that employers will try to evade incentives created by this kind of disability program reform.

Disability Insurance in the Netherlands

Since 1967, the Disability Insurance program in the Netherlands has been provided as a public scheme that is mandatory for all workers. Disability benefits are provided if workers experience a loss of income capacity due to medical impairments of 35 percent or more.[1] For these workers, benefits provide insurance for 70 percent of the loss of income due to impairments. Since 2004, workers can apply for disability benefits after two years of sickness. During the so-called "waiting period," employers are responsible for the provision of reintegration activities (services and/or adaptations that facilitate the worker's return to work), and for the continued payment of wages. Disability insurance claims are assessed and premiums are set by the public social benefit administration called the UWV (*Uitvoeringsinstituut Werknemersverzekeringen*), which roughly translates as Employee Insurance Agency. The UWV determines the presence of impairments, the consequences for the earnings potential of an applicant, the degree of disability as a percentage of the worker's former wage, and the corresponding disability benefit level. Workers may thus receive benefits for partial disability, which are supplemented by unemployment insurance benefits—and subsequently by social assistance benefits—if the residual earnings potential is not used sufficiently. Figure 2 shows that in 2013, 71 percent of all disability benefit recipients were classified as 100 percent disabled and thus received full disability benefits, whereas workers with 15 to 35 percent loss of earning capacity—constituting 10 percent of all recipients—were the second-largest

[1] For workers with residual capacities, a set of regular jobs is selected that meet the worker's physical and mental impairments. Based on the wage rates of these jobs, the residual earnings capacity is determined.

Figure 2
Shares of Partially and Fully Disabled DI Benefit Recipients in the Netherlands, 2013



Source: UWV (2012).

group. It should be noted that the threshold value for disability benefits was raised from 15 to 35 percent in 2006. Thus, workers with a disability degree that is lower than 35 percent entered the scheme before 2006.

The Dutch Disability Insurance system has two important institutional features that differ from systems in most other high-income countries. These features haven't changed much since inception in 1967, not even after the reforms that started in the late 1990s. First, the Dutch disability program covers all workers against all income losses that result from both occupational and nonoccupational injuries. In most other high-income countries, eligibility for disability insurance is constrained by work history requirements or limited to occupational injuries only. Including all workers against the whole gamut of medical contingencies increases the possibility of sizable screening errors in disability determinations (as discussed, for example, in Parsons 1991) where the social benefit administration is more likely to prioritize on minimizing erroneous denials ("Type I errors") at the cost of increasing erroneous admissions ("Type 2 errors"). Clearly, the sharp rise of disability enrollment in the Netherlands for two decades after the mid-1960s and the continued high levels of disability for two decades after that suggests that applicants have in the past successfully exploited this feature of the Disability Insurance system (Burkhauser, Daly, and de Jong 2008).

Second, wage payments for sick workers are continued in the waiting period that precedes disability claims. This scheme was funded from sectoral insurance premiums until 1996, and privatized since then—that is, employers are responsible for the continued payment of wages during sickness. Clearly, the Dutch sickness benefit scheme differs from the US system, in which individuals who are typically no longer working and receiving wages must take the initiative to submit disability applications. With continued wage payments during a period of sickness and prior to the disability assessment, the Dutch system does not provide strong incentives for disabled (or sick) workers to resume work quickly. As a result, workers with less-severe health problems are less likely to screen themselves out of DI benefit receipt.

These two main institutional features of Dutch Disability Insurance, with their incentives for broad coverage of impairments and limited self-screening, thus laid the ground for high Disability Insurance inflow rates after the program's inception in 1967, and, accordingly, a continuous increase of DI enrollment to unprecedented levels.[2] The relative attractiveness of disability vis-à-vis unemployment insurance effectively triggered workers and employers to take advantage of the scheme. Compared to unemployment insurance, which also covers 70 percent of the loss of income, the Dutch disability system provides benefits with entitlement periods that are unrestricted and without the job search requirements that apply to the unemployed. Moreover, statutory disability benefits were (and are) often supplemented by nonstatutory benefits for specific collective labor agreements, raising the replacement rate of workers from 70 to 80 or even 90 percent in the first years of receiving disability benefits (van Vuren and van Vuuren 2007).

In the past, moving unwanted workers into Disability Insurance rather than into unemployment insurance has also been attractive for employers. Until 1996, employers did not bear the costs of sick pay and Disability Insurance benefits for their own employees. However, if employers fired a worker, especially an older worker with a long work history, the employer faced substantial costs. In the Netherlands, the general rule applies that each additional year of working history implies one extra monthly salary as severance pay. For older workers, this means that the amount of severance pay could be equal to three to four years of annual salary. As a result, many employers preferred to use disability insurance as a substitute pathway for unemployment insurance, even if there was the risk that the disability claim would not be awarded at the end of the sickness benefit period. De Jong (2008) concludes that the disability insurance scheme has been used in this way to support the transformation from an industrial to a service-oriented economy by facilitating massive lay-offs in vulnerable sectors. For many workers in these sectors, disability effectively functioned as an early retirement route (Kerkhofs, Lindeboom, and Theeuwes 1999). Thus, workers and employers had a mutual interest in using the disability insurance scheme as a substitute pathway into unemployment and early retirement.

Although the potential for substitution effects between disability and unemployment is self-evident, inferring the actual size of hidden unemployment within

[2] While Dutch Disability Insurance inflow rates have varied substantially over time, DI outflow rates have been fairly constant over time, ranging around 11 percent. As a result, only limited variation in DI enrollment rates can be explained by (variation in) DI outflow rates.

those categorized as disabled is not an easy task. When workers have become incapable of performing their current tasks, either medical or functional criteria may predominate. This renders it almost impossible to know if an individual is "hidden unemployed"—particularly when someone has entered into the disability insurance scheme only recently and the person's remaining work opportunities are, as yet, undiscovered.

To circumvent these problems, studies of the importance of substitution effects between disability and unemployment typically rely on indirect inferences based on inflow rates to disability in a given year, or on overall disability enrollment rates, to assess the overall size of hidden unemployment. For example, Autor and Duggan (2006) point out that application rates for disability insurance are countercyclical—that is, application rates for disability rise during recessions—while illness is not itself directly countercyclical, which suggests substitution effects between disability and unemployment insurance. Koning and van Vuuren (2007, 2010) follow a similar research strategy for the Netherlands, seeking to explain inflow rates for disability and unemployment insurance. Without substitution effects between the two insurance programs, average wages and sectoral growth levels should affect only the numbers of those receiving unemployment insurance and not the numbers receiving disability. However, both these variables do affect inflows to disability benefit receipt, and in this way Koning and van Vuuren infer that about one-quarter of the inflow into disability insurance from 1993 to 2002 consisted of hidden unemployment. Aarts and de Jong (1992) take an alternative approach. Using medical information of disability benefit recipients in the 1980s, they find that hidden unemployment among recipients of disability insurance benefits ranges between 33 and 51 percent.

Assessing the Effectiveness of Disability Policy Reforms

Policymakers in the Netherlands started reforming Disability Insurance in the early 1990s, and while these efforts at first seemed promising, these efforts did not persist. For example, disability benefits were reduced in 1993; these declines were largely offset by (almost) equal increases in supplementary private benefits in the following years, leaving the payments to those who were disabled much the same. Another step in 1993 was the start of a large-scale program of medical re-examinations of existing recipients of disability benefits. These re-examinations had a large effect, amounting to a decline in the probability of receiving disability benefits of 5 percentage points (Borghans, Gielen, and Luttmer 2014). (About 30 percent of the reduction in disability insurance spending was cancelled out by additional spending on unemployment insurance and social assistance.) But these measures were not politically sustainable and ended after two years.

However, in the following years, the Dutch government implemented reforms that persisted and substantially affected disability inflow rates. We will cluster these reforms in three broad categories: 1) enhancing employer incentives to avoid disability insurance; 2) increasing screening for disability; and 3) tightening eligibility for continued receipt of disability benefits and increasing work incentives for recipients.

Enhancing Employer Incentives (1996, 1998)

Starting in 1996, the Dutch government undertook a series of policies to change the incentives of employers so they would be less eager to facilitate the movement of workers to disability. The idea was that employers should be made responsible for a substantial part of the sickness and disability benefit costs of their workers, thus encouraging activities that would prevent sickness and disability and reintegrate the sick and disabled into the workforce. To start with, the sickness benefit program was privatized in 1996, making employers fully responsible for these costs. Employers could reinsure this risk with private insurers or bear this risk themselves. This change in the program resulted in a decline in absence rates (De Jong and Lindeboom 2004).

In 1998, the disability insurance system was experience-rated: that is, the amount that employers pay into Disability Insurance was linked to the employers' past experience of employees receiving disability. Specifically, employers were to bear the costs of the first five years of Disability Insurance benefits. (In 2006, this experience-rating period was extended to ten years.) Initially, the experience-rating system did not cause substantial controversy among employers and policymakers. By 2003, the experience-rating incentive had reached its maximum impact, and about 31 percent of all disability insurance costs were experience-rated (Koning 2009).

Given that the privatization of sickness pay and the introduction of experience rating for disability insurance were the key policy reforms that were taken between 1996 and 2001, one might conclude that this alteration in employer incentives did not make a substantial difference. After all, looking back at Figure 1, the inflow rates to receipt of disability benefits varied from 1.1 to 1.4 percent of the insured population between 1996 and 2001, which is only a little lower than between 1990 and 1995. This simple eyeball test would thus suggest that even with a change in incentives, employers had limited ability to prevent inflows to the disability rolls.

However, there are strong reasons to believe that the effectiveness of experience rating increased in later years. Koning (2009) argues that the effects took substantial time to come into force, in part because many employers were initially unaware of the details of the new system. Particularly from the perspective of small and medium-sized firms, the experience-rating system was complex, and it was seemingly unimportant—as long as employers were not seeing a close connection between flows from their company into disability benefits and their employer-paid disability insurance premiums. Along similar lines, Hyatt and Thomason (1998) and others have argued that the awareness of experience rating among individual firms may be limited. Moreover, employer awareness of experience rating seemed especially low for firms who were benefitting from lower Disability Insurance premiums.

To estimate the potential effect of experience rating, Koning (2009) employs a difference-in-differences strategy, looking at changes in registered inflow rates to disability benefits for employers who *had*, versus *had not*, experienced premium raises (so far). This strategy takes advantage of the rule that past inflow into disability benefits affected premiums with a lag of two years (as such, mean-reversion effects are controlled for). Following this approach, the response to an unanticipated increase in disability insurance premiums is estimated to be a 15 percent decrease in the disability inflow rate. The experience-rating plan was thus effective for individual employers, but its macro-effect had to accumulate over time. Many employers still needed a "wake-up call" to pay attention to experience rating and subsequently increase activities that could prevent future sickness and disability.

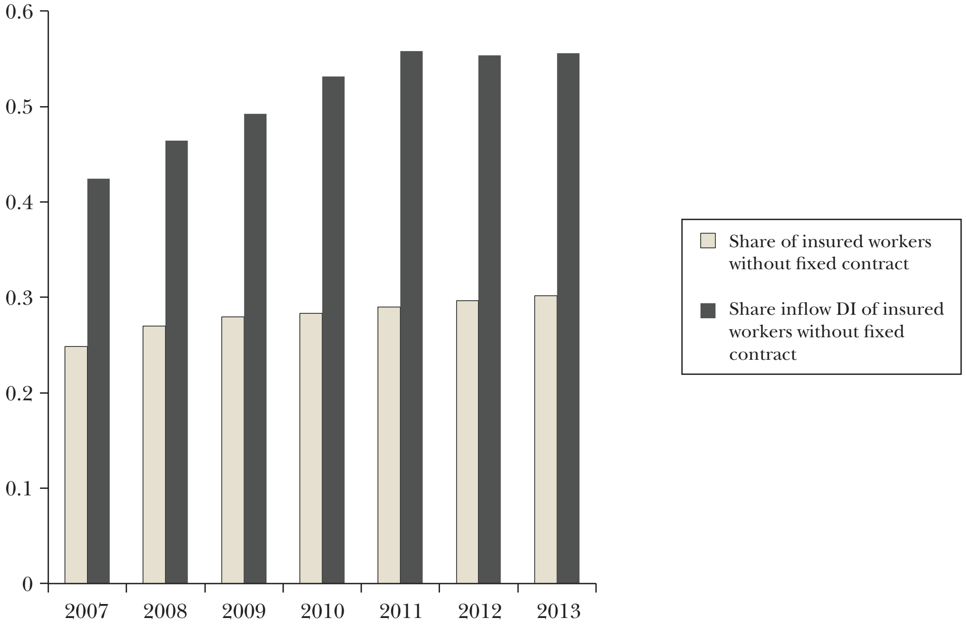
While awareness of the experience-rating plan among employers has grown, criticism of experience rating has grown as well. This is not surprising, as the Netherlands stands out as the country with probably the highest experience-rating incentives relating to disability insurance in the world today. Employer organizations argue that they cannot bear the financial risks associated with experience rating, which, after all, are added to the sick-pay costs that were already there during the waiting period and also cover disability for nonoccupational reasons.[3] Moreover, Dutch employers typically have no room to appeal the decision to award disability benefits (in the context of workers' compensation claim decisions, there is usually room to appeal, as discussed in Tompa, Cullen, and McLeod 2012).

In this context, the most straightforward way for Dutch firms to circumvent experience-rating incentives in the Netherlands is to hire workers with temporary contracts. These individuals are sometimes labeled as "safety netters." If temporary and flexible workers are awarded disability benefits, the costs are not assigned to individual employers but financed by collective funds. Thus, one would expect to see an increase in the share of temporary or flexible employment, particularly of high-risk workers with bad health conditions.

Although there is no causal evidence on the effect of experience rating on type of labor contract offered, a basic comparison of the rate of inflow to disability from workers with fixed and temporary contracts suggests that sorting effects have become more important. In particular, Figure 3 shows that the share of disability benefit awards to "safety netters" out of the total number of disability awards has increased from 42 percent in 2007 to 55 percent in 2011 (UWV 2013). This trend cannot be entirely explained by the (much smaller) decrease in the share of workers with permanent contracts; it rather suggests that vulnerable groups with bad health conditions have sorted into flexible jobs. Thus, although employers are not allowed to screen out workers with health conditions when doing permanent hires, one might doubt the enforceability of this law. This pattern raises concerns about the success of experience rating as well as the notion that employers should play a key

[3] To illustrate the implications of wage continuation and experience rating, suppose a worker becomes fully disabled; this means that the employer can become responsible for two years of full wages for this worker along with ten years of disability benefits.

Figure 3
Workers in Flexible and Temporary Jobs Expressed as a Share of the Total Number of Insured and as a Share of DI Inflow (2007–2013)



Source: UWV (2013b).

role in a program for the well-being of workers. We will return to this issue in the next sections.

Stricter Screening: The Gatekeeper Protocol (2002)

The introduction of the Gatekeeper protocol in 2002 is generally considered to be the most effective policy measure that has been taken to curb the rate of those receiving disability benefits. The Gatekeeper protocol specifies the legal responsibilities of both the employer and the incapacitated worker during the period of sickness and absence before the worker applies for disability benefits. The protocol means that the social benefit administration (the UWV) is no longer involved in the process of reintegrating sick workers during the waiting period but acts purely as a gatekeeper.

The Gatekeeper protocol spells out the required behavior of employers and workers starting with the first weeks of absence from the job. In particular, after a maximum of six weeks of absence, the employer and worker should make a first assessment of medical cause and functional limitations. Based upon this assessment, they subsequently must draft a return-to-work plan within eight weeks of absence. This plan should include several dates to evaluate and modify the plan, if relevant. If the worker has not fully returned to work at the end of the waiting period, the worker then files a disability benefit claim. Benefit claims are only considered admissible by the social benefit administration if they are accompanied by a return-to-work report, containing the original plan and an assessment as to why the plan has not (yet) resulted in work resumption. If the procedure was not followed, the employer may be obliged to continue providing sick pay for some additional months rather than having the worker transfer to disability benefits.[4]

In this way, the Gatekeeper protocol encourages the disability prevention and reintegration activities of employers. The protocol forces employers to focus their attention at the onset of sickness, when the opportunities for recovery and work resumption are probably most substantial. The stricter screening also triggers mechanisms of self-selection and self-screening among applicants with less-severe health conditions (Parsons 1991). So the protocol with its stricter screening involves stronger incentives—both for employers and workers.

The Gatekeeper protocol appears to have had an immediate impact on the behavior of employers and workers. For example, Figure 1, presented earlier, showed a sharp decrease in the percentage of the population receiving new disability insurance awards, from 1.4 percent of the insured population beginning disability benefits in 2001 to 0.8 percent in 2004. The Gatekeeper protocol was the only reform that took place during this time.

Using quarterly data, van Sonsbeek and Gradus (2013) investigate the contribution of the Gatekeeper protocol and some other measures on the decrease in disability inflow rates. They argue that these policies have reduced the disability award rates by about 40 percent, compared to the level prior to 2002. As this effect is far more substantial than the (immediate) impact of employer incentives, one could well argue that the Gatekeeper protocol has made the costs of wage continuation and experience rating more salient to employers.

To shed more light on the mechanisms explaining the reduced inflow levels, de Jong, Lindeboom, and van der Klaauw (2011) exploit a field experiment with regional variation of the intensity of screening by the social benefit administration. According to their analysis, stricter screening causes both self-selection and increased effort to resume work during sickness absenteeism—with both effects of about equal size. These mechanisms seem to have been strengthened when the mandatory waiting period of absence before receiving disability benefits was extended from one to two years in 2004. In our view, this extension of the mandatory waiting period is best understood not as a separate reform, but as part of the Gatekeeper protocol.

Although the Gatekeeper protocol seems to have contributed to the decrease in inflow rates to disability, there is concern that it may have had some unintended

[4] So far, the number of lawsuits that occur when employers and workers disagree on a return-to-work plan, or when plans are not executed, is limited. There are various reasons for this. First, employers face the risk of continued wage payments if there is no return-to-work plan. Second, workers can get fired by their employers if they do not cooperate. Third, mediators from the public employee insurance agency can be contacted in case of disagreements. Ultimately, if the employer and worker still disagree after this, a lawsuit may well occur.

effects. First, one concern is that workers with less-severe health conditions have sorted into other social benefit schemes, with unemployment insurance as the most likely candidate. However, de Jong, Lindeboom, and van der Klaauw (2011) find no evidence that increased gatekeeping by UWV resulted in more inflow into unemployment insurance, suggesting that most workers who did not receive disability benefits under the Gatekeeper program resumed their work. Second, and similar to the enhancement of employer incentives prior to 2002, the protocol might have made employers more hesitant to hire workers who have a higher risk of bad health. We return to this issue at the end of this article.

2006: Tightening Eligibility Criteria and Increasing Work Incentives

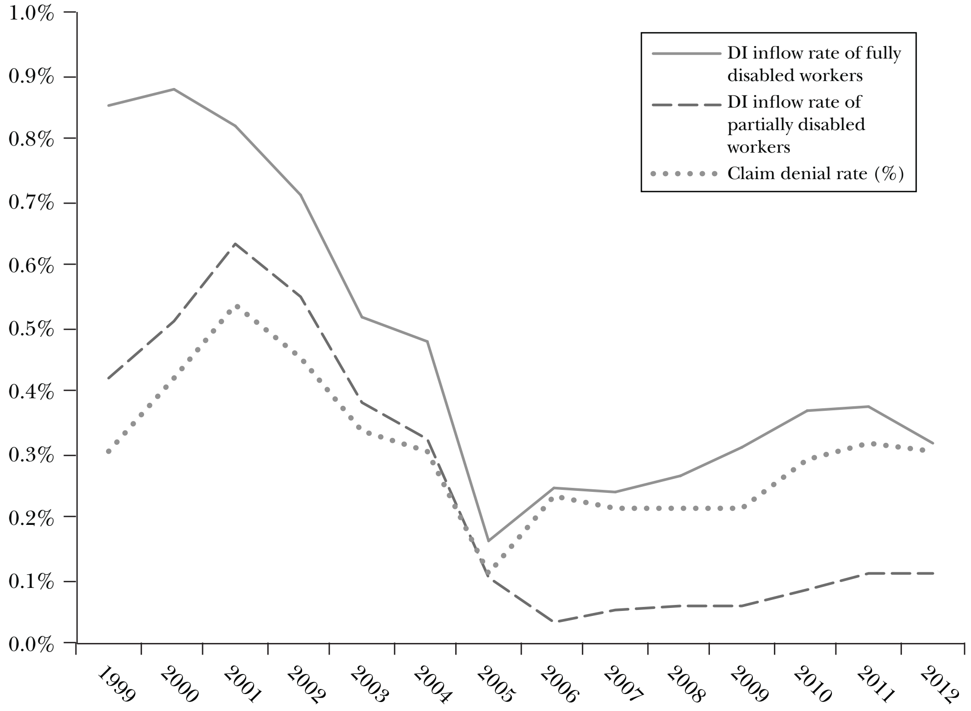
The most recent disability insurance reforms entailed the replacement of the old Invalidity Insurance Act ("WAO" or *Wet op de Arbeidsongeschiktheidsverzekering*) by a new disability law called the Work and Income (Employment Capacity) Act that included new benefit conditions ("WIA" or *Wet Werk en Inkomen naar Arbeidsvermogen*). Although there was a widespread belief that the previous, inflow-related policy measures were effective in curbing inflow to disability benefits, policymakers felt that the program still was not effective in assisting disabled workers in reaching their full employment potential. The rates of recovery and work resumption for disabled workers were still negligible—although many of the impairments had been expected to be temporary. Therefore, to stimulate the work resumption of workers—particularly those with temporary and less-severe impairments—the new disability law included three major changes.

First, the new disability insurance program introduced the distinction between two types of benefits: one for workers who are fully and permanently disabled and one for workers who are partially and/or temporary disabled. For the group of fully and permanently disabled, disability benefits were raised to 75 percent of the last earned wage. Admission to this scheme has been very strict and limited to a selective group of impairments that are expected to be permanent. Consequently, the yearly inflow rate is only about 0.1 percent of the insured working population. The idea behind this distinction was that the room for moral hazard would be negligible for the small group of workers with severe and permanent impairments. Consequently, benefit levels could be increased and employers were no longer held financially responsible for this group.

Second, the eligibility criteria for the partial scheme have been tightened by raising the minimum degree of disability from 15 to 35 percent of the previously earned wage. Workers with less-severe health impairments are thus expected to continue their employment with some adaptations or—if they are fired—to apply for unemployment insurance. Figure 4 shows that this way of tightening Disability Insurance eligibility has led to a sharp increase in the number of claim denials since 2006. This is mirrored by a decrease in the inflow into partial DI schemes that are awarded, while the inflow into full DI schemes has increased to some extent. In this respect, van Sonsbeek and Gradus (2013) argue that the higher disability threshold decreased disability insurance award rates by about 25 percentage points. Since its

Figure 4

Inflow Rate into Disability Insurance and the Annual Number of Claim Denials, 1999–2012



Source: UWV (2012).

Note: Here "fully disabled" is defined as a degree of disability higher than or equal to 80 percent, and "partially disabled" refers to workers with a disability degree below 80 percent.

inception in 2006, the reform is thus changing the composition of workers receiving disability benefits. By increasing the degree-of-disability thresholds in awarding benefits, the Netherlands system moves closer to that of other OECD countries, most of which have substantial thresholds.

Third, the new system introduced wage subsidies to encourage partially disabled workers to use their remaining earnings potential. Similar to the system before 2006, partially disabled individuals receive wage-related benefits that replace 70 percent of the difference between their pre-disability wage and their wage potential in the first years of their benefit. The length of this period is determined by their working history and lasts 38 months at maximum. After this period, however, workers only continue receiving this level of disability benefits if they work more than 50 percent of their residual earnings capacity. Otherwise, their benefit level is set equal to the level of social assistance.

There are strong reasons to believe that the introduction of the wage subsidy for partially disabled has had only a limited impact. Since 2006, only 29 percent of the disability awards consist of workers that are diagnosed as partially disabled. In addition, many collective bargaining agreements have provided nonstatutory benefits to offset the drop in statutory benefits in the follow-up period. One also should keep in mind that the targeted individuals have been out of the workforce for several years—starting with the waiting period of sick pay of two years and followed by some years of benefits that are wage-related and do not inhibit strong work incentives. Similar to the experiences with the US Ticket-to-Work program, which also seeks to encourage the disabled to return to work, it is likely that the readiness to resume work has eroded during the period away from the workforce (Autor 2011).

Summing Up the Dutch Reforms

The key to the success of disability insurance reform in the Netherlands has been the intensified role of employers in preventing long-term sickness, absence, and subsequent disability, with a strong emphasis on early interventions. The employer incentives increased the economic urgency among employers to exert sickness and accident prevention and workforce reintegration activities, while the Gatekeeper protocol has facilitated employer awareness and guided employers in their new role. Most of the gains in curbing inflow to disability benefits have been made in the waiting period that precedes the application of claims.

The new disability law that started in 2006 has made a smaller but still substantial contribution to the decreased inflow to disability benefits. The main effect came from a tightening of eligibility criteria, which caused fewer partially disabled workers to be awarded disability benefits. This probably has limited the ability of the new system to provide well-targeted and effective return-to-work incentives to the less-severely disabled.

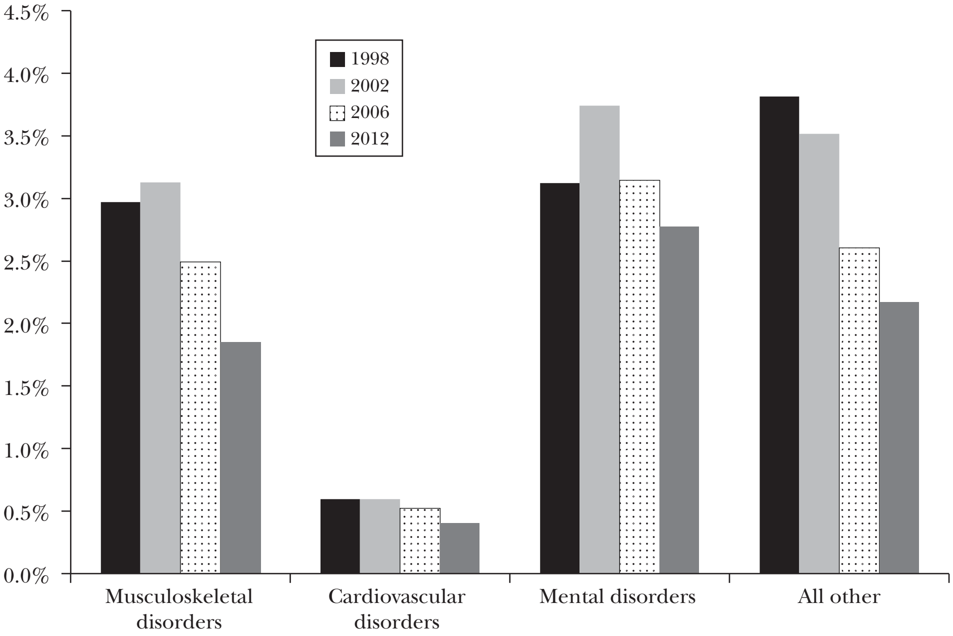
One major concern with the reforms is the high level of obligations and financial risks born by employers. As a consequence, employers may now be more reluctant to hire vulnerable workers, in particular those with existing health conditions. In what follows, we will therefore consider the position of vulnerable worker groups: how did the reforms impact the structure of impairments that end up receiving disability benefits, and how have the employment probabilities of disabled workers evolved over time?

Labor Market Effects among Disabled Workers

Which Impairments Were Affected Most by the Reforms?

With reforms that focused on enhancing the screening for less-severe impairments and encouraging re-entry to the workforce, one would expect major shifts in the composition of disability beneficiaries. Figure 5 shows the evolution of the percentage of awards by diagnosis groups. The reforms seem to have affected all broad impairment types, but to different degrees. The percentage of Disability Insurance recipients with musculoskeletal disorders per insured has almost halved

Figure 5
Disability Insurance Enrollment Rate per Insured Worker, Stratified by Diagnosis Group (1998, 2002, 2006, and 2012)



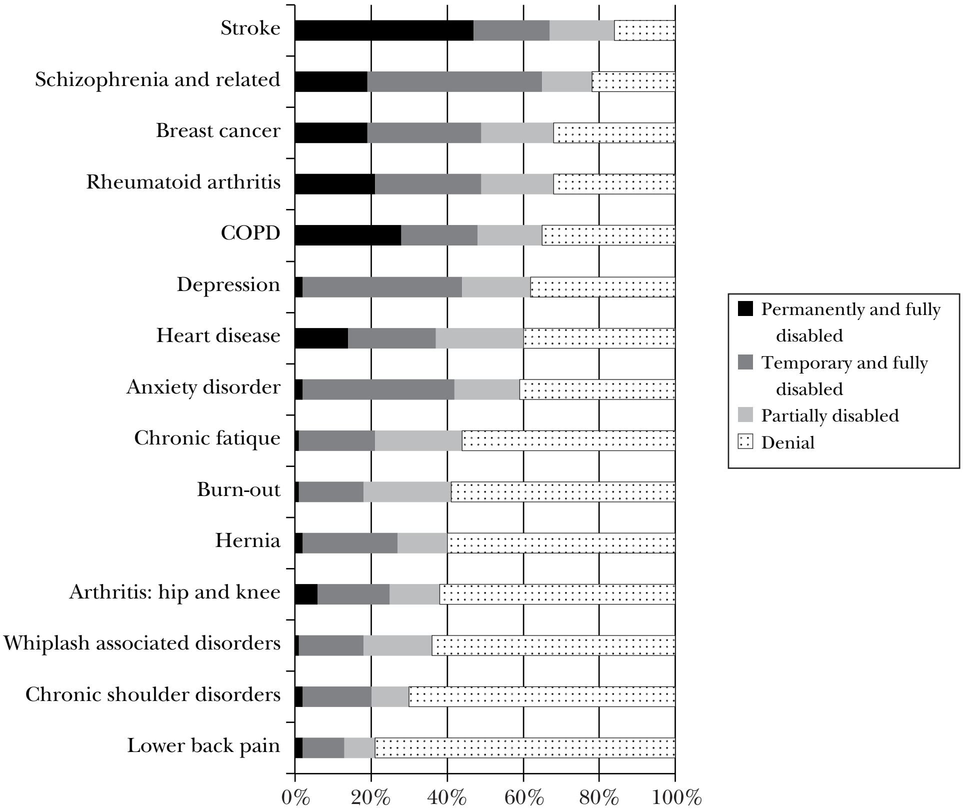
Source: UWV (2012).

since 2002. This dramatic decrease largely coincides with a more general decrease in findings of partial disability. Individuals with less-severe impairments—for example, those with modest levels of lower back pain—have either resumed work in the two-year waiting period of sickness or did not meet the criteria of the new disability scheme.

The decrease in disability in the Netherlands has been accompanied by only a small reduction in the rate of disability awards due to mental disorders. Indeed, mental disorders made up 29.7 percent of the diagnoses for new disability enrollment in 1998 but were 38.5 percent of the new diagnoses in 2012 (UWV 2012). This greater relative importance of mental disorders as a cause of disability is a trend that most OECD countries are facing, with some countries—like Sweden and Denmark—having even steeper increases in the share of disability awards due to mental disorders (OECD 2011). The high incidence of mental disorders among the disabled helps to explain why it has proven difficult to bring disabled workers back into the workforce. Those in this category are often labeled as "fully and temporarily" disabled, but in practice, the number of workers in this category that fully recover has proven to be negligible, and many of these individuals will eventually

Figure 6

Distribution of Most Important Diagnosis Groups across Disability Insurance Benefit Types and Application Denials in 2006



Source: UWV (2007).
Note: COPD is chronic obstructive pulmonary disease.

transfer to the more generous scheme for permanently disabled individuals, rather than back to work.[5]

Figure 6 takes a closer look at the types of diagnoses that are made for disability applications in 2006, the first year of the new disability scheme. Three patterns in these data are worth noticing. First, almost all disability applications with musculoskeletal disorders as the primary impairment—that is, lower back pain, chronic shoulder disorders, and hernia—are denied and virtually have no chance of being

[5] Of the temporary and fully disabled workers that entered the disability insurance system between 2006 and 2010, only 14 percent have left the scheme (de Jong, Everhardt, and Schrijvershof 2013). These exits from the status of temporary and full disability include those who reach retirement age and those who are reclassified as permanently and fully disabled workers, along with those who have at least a partial recovery.

awarded full and permanent benefits. Second, fully and temporarily disabled workers are an important group among the more severe mental disorders, like schizophrenia, depression, and anxiety disorders. Finally, only a few severe impairments—such as stroke and chronic obstructive pulmonary disease (COPD)—have a substantial probability of being qualified as fully and permanently disabled. Once more, this reflects the stringency of the new system.

When taking a broader perspective, Figure 6 also reveals that the largest share of benefits awarded are effectively experience-rated; it is only for workers with the most severe impairments that DI benefits costs are not borne by the employer. As the risk of these impairments is probably outside the control of the employer, this way of differentiating seems likely to be efficient. At the same time, one would expect the degree of experience rating to be highest for impairments that are related to work, particularly for musculoskeletal disorders. But since these physical impairments have the lowest probability of being awarded with benefits, the effective preventative impact of experience rating will be limited for this group.

Labor Supply Effects

The changes in the Dutch disability system, and in particular the changes of 2006, aimed at stimulating work resumption rates of those with temporary and less-severe impairments. As disability enrollment rates have declined dramatically, the natural question that arises is whether these changes are accompanied by increases in employment rates of those with impairments relative to their healthy counterparts. To explore this issue, we use the POLS health survey from Statistics Netherlands (*Permanent Onderzoek Leefsituatie*) to describe trends in employment rates of those in good health versus those in bad health.[6] The share of individuals between 25 and 65 years of age that report bad health is fairly constant around 20 percent in all years in our sample.

Table 1 depicts differences in employment rates between individuals in bad and good health—labeled the "health employment gap"—for both men and women in the POLS data. These employment gaps can be substantial, ranging from 20 to about 30 percentage points of the sample. The figure shows for males a reduction in the health employment gap of about 5 percentage points since 2002. For females, the gap remains more or less constant over time. It should be noted that the different trend for females is in part due to increases in participation rates among healthy females as well. For men, however, participation rates among healthy individuals are almost constant over time.

With declining employment gaps for males, it becomes relevant to see whether this is reflected in differences in benefit receipt *of any benefits* between those in good

[6] The data are the *Permanente Onderzoek Leefsituatie* (POLS) data from 1998, 2002, and 2006 and the *Gezondheidsenquête* Health Survey of 2010. The POLS data consist of repeated cross-sections and come with sample weights that we use to construct our figures. Bad health is derived from the response to a question regarding an individual's general health and equals one if the response is fair, bad, or very bad. Good health is defined as the complement of bad health (corresponding to a response of good or very good). Employment is defined as having a paid job and working more than 12 hours per week.

Table 1

Health and Employment of Males and Females (1998, 2002, 2006, and 2010)

	1998	*2002*	*2006*	*2010*
Males				
Employment rate of individuals with good health (%)	86.1	86.7	84.9	83.4
Employment rate of individuals with bad health (%)	57.6	55.8	59.0	57.2
Health employment gap (%)	**28.5**	**30.9**	**25.9**	**26.2**
Females				
Employment rate of individuals with good health (%)	56.4	65.5	68.9	71.0
Employment rate of individuals with bad health (%)	35.1	35.7	40.6	40.0
Health employment gap (%)	**21.3**	**29.8**	**28.3**	**31.0**

Source: POLS.
Note: The health employment gap is defined as the difference between the employment rates of individuals with good health and bad health, measured in percentage points.

health and bad health. Here, "benefit receipt" broadly includes disability benefits, unemployment benefits, social assistance (for those with low incomes) and early retirement benefits. To shed more light on this, Table 2 shows for males a drop of 10 percentage points from 2002 to 2010. While there has been a general decline in benefit recipient rates of all men, the decline in benefit rates of those in poor health was of course considerably stronger. The 10 percentage point drop in the benefit gap between unhealthy and healthy men is larger than the about 5 percentage point reduction in the employment gap, implying that some of those who have left benefits did not obtain "substantive gainful employment." For women, one can observe a slight increase in the benefit receipt of those in bad health versus those in good health over the longer period from 1998 to 2010. Again, this may well stem from increases in participation rates of women in good health as well.

These descriptive analyses suggest that the Dutch Disability Insurance reforms probably enhanced the work continuation of male individuals with poor health. At the same time, however, the share of less-healthy males without work and receiving no benefits has increased as well. This finding could imply that some disabled people who are unable to work are being rejected for disability insurance, or that it has become harder for marginally healthy workers to claim disability benefits, or both.

Discussion and Outlook

The key to the success of Disability Insurance reform in the Netherlands has been the intensified role of employers in preventing long-term sickness, absence, and subsequent inflow to receipt of disability insurance benefits. The Gatekeeper protocol implemented in 2002 has provided employers guidance to implement

Table 2
Health and Benefit Receipt of Males and Females (1998, 2002, and 2010)

	1998	*2002*	*2010*
Males			
Benefit rate for individuals with good health (%)	9.4	7.1	8.3
Benefit rate for individuals with bad health (%)	47.8	46.9	38.1
Health benefit gap (%)	**38.4**	**39.8**	**29.8**
Females			
Benefit rate individuals with good health (%)	9.8	8.4	7.7
Benefit rate individuals with bad health (%)	34.4	37.6	36.9
Health benefit gap (%)	**24.6**	**29.2**	**29.2**

Source: POLS.
Notes: The health benefit gap is defined as the difference between the benefit rates of individuals with bad health and good health, measured in percentage points. Here, "benefit" broadly includes disability benefits, unemployment benefits, social assistance (for those with low incomes), and early retirement benefits.

their new role, while the tighter eligibility criteria since 2006 seem to make the Dutch disability system less susceptible to providing disability benefits to those who can still work. These reforms probably have improved the targeting efficiency of the Dutch disability system, leading to higher employment rates among male individuals with both bad and good health. At the same time, however, the number of marginally healthy workers without benefits and without work has increased as well.

Given the decline in the rates at which disability benefits are being awarded each year, one might be inclined to think that the rates of people receiving disability as a share of the workforce will decline further in the years that come. However, the Dutch Disability Insurance system still includes some features that may undermine its long-term sustainability. In what follows, we will discuss two features that may have relevance for many other high-income countries: increased labor market flexibility and the inability of the program to get disabled workers to resume work—even for those whose impairments are temporary.

Changing Employer Incentives

The cornerstone of the current Dutch disability insurance system is the interest that employers should have in investing in the health and safety of their workers. However, this interest implicitly assumes long-standing or near-permanent employment contracts. In this context, some health problems accumulate over time and investments in workplace health and safety may take time to effectuate. With a continuous rise of flexible and temporary contracts, the case for sick pay costs and experience rating that stretch out over a long time window becomes weaker. Indeed, some argue that the financial risks of sickness and disability are too high for some firms, reducing the flexibility they need to adapt to labor market conditions.

As we argued earlier, it is likely that employers have responded to the incentives by hiring high-risk workers on a temporary basis only. One obvious policy response would be the introduction of employer-incentives-linked disability benefits for those hired on temporary and flexible contracts as well. Recently, the Dutch government decided to implement such plans beginning 2016: that is, employers will also be responsible for workers on temporary and flexible contracts both during the two-year waiting period of sick leave prior to the disability application and, given experience rating, after inflow into the disability system. This change will discourage substitution into temporary contracts, but serious doubts could be raised about the ability of employers to influence the risk of sickness and disability for many short-term and temporary workers.

Policymakers currently are also considering other ways of redesigning employer incentives to curb sickness and disability, searching for methods that would provide strong incentives but with lower financial risks for employers. The most likely candidate appears to be the adjustment of the experience rating of disability insurance premiums by somehow reducing the share of disability costs that would be passed back to the employer. Whether this could be done by means of a shorter time window for experience rating or a lower percentage payment is a question that calls for further research into the optimal design of incentives in this system.

Activating Disabled Workers

While the Dutch disability reforms have been successful in curbing inflow rates to disability benefits, the system has become less effective in enhancing and employing the residual work capacity of workers that are awarded benefits. This difficulty is not surprising, because a substantial group of workers with less-severe impairments is no longer eligible for disability benefits. Moreover, many of those diagnosed as temporarily and fully disabled are mentally impaired workers with low education levels (de Jong, Everhardt, and Schrijvershof 2013). Even in cases where one might expect these individuals to improve for medical reasons, the switch to substantial and gainful employment is only rarely observed.

One contributing explanation for the persistence of this area of disability can be found in the design of work incentives in the new scheme. At present, a fully disabled worker who finds partial employment will then have his or her level of disability reassessed. In addition, the switch from full to partial disability incurs the risk of not finding employment with sufficient earnings in order to receive the wage supplement. In effect, the current setup of incentives thus effectively encourages fully and temporary disabled workers to abstain from work—even if their health recovers in a way that would allow them to regain part of their earnings potential. As we pointed out earlier, a related problem is that the incentive to return to work often arrives too late to make a difference, given how long individuals have already been out of the labor force. When putting this in a broader perspective, the question arises whether the Dutch Disability Insurance program puts too much emphasis on employer incentives while the effectiveness of worker incentives is still limited.

Closing Remarks

Workers with poor health and low productivity levels are a vulnerable group in the labor market and will pose a challenge for policymakers in any country that provides disability benefits. Because the definition of disability depends explicitly on the job opportunities of workers, there always will be beneficiaries of disability payments who are capable of working but insufficiently productive to earn their own living (see also Autor and Duggan 2006). This particularly holds for countries where statutory minimum wages are relatively high—such as the Netherlands, as well as Sweden and Norway—resulting in more limited opportunities for (formerly) disabled workers to resume work. There is inevitably a group of workers for whom early interventions do not hold much promise and for whom working is not viable—whether these workers are classified as disabled or not.

■ *David Autor, Chang-Tai Hsieh, and Timothy Taylor are gratefully acknowledged for their constructive comments to this paper. Also, we would like to thank Carla van Deursen from the Dutch Employee Insurance Agency (UWV) for providing us with some of the statistical information that we used. Pierre Koning declares that he is (part-time) employed as Chief Science Officer at the Ministry of Social Affairs and Employment in the Netherlands.*

References

Aarts, Leo J. M., and Philip R. De Jong. 1992. *Economic Aspects of Disability Behavior.* Elsevier Science.

Autor, David H. 2011. "The Unsustainable Rise of the Disability Rolls in the United States: Causes, Consequences, and Policy Options." NBER Working Paper 17697.

Autor, David H., and Mark G. Duggan. 2006. "The Growth in the Social Security Disability Rolls: A Fiscal Crisis Unfolding." *Journal of Economic Perspectives* 20(3): 71–96.

Borghans, Lex, Anne C. Gielen, and Erzo F. Luttmer. 2014. "Social Support Substitution and the Earnings Rebound: Evidence from a Regression Discontinuity in Disability Insurance Reform." *American Economic Journal: Economic Policy* 6(4): 34–70.

Burkhauser, Richard V., Mary C. Daly, and Philip R. de Jong. 2008. "Curing the Dutch Disease: Lessons for the United States Disability Policy." Michigan Retirement Research Center Working Paper 2088-188, University of Michigan.

Hyatt, Doug, and Terry Thomason. 1998. *Evidence on the Efficacy of Experience Rating in British Columbia. A Report to The Royal Commission on Workers' Compensation in BC.* http://www.qp.gov.bc.ca/rcwc/research/hyatt-thomason-experience.pdf.

Jong, Philip R. de. 2008. "Recent Changes in Dutch Disability Policy." September 2008. http://www.ape.nl/include/downloadFile.asp?id=75.

Jong, Philip de, Tom Everhardt, and Carlien Schrijvershof. 2013. "Duurzaam niet-duurzaam? Onderzoek naar niet-duurzaam volledig arbeidsongeschikt verklaarden." APE-report nr. 967, APE, February.

Jong, Philip de, and Maarten Lindeboom. 2004. "Privatisation of Sickness Insurance: Evidence from the Netherlands." *Swedish Economic Policy Review* 11: 11–33.

de Jong, Philip, Maarten Lindeboom, and Bas van der Klaauw. 2011. "Screening Disability Insurance Applications." *Journal of the European Economic Association* 9(1): 106–29.

Kerkhofs, Marcel, Maarten Lindeboom, and Jules Theeuwes. 1999. "Retirement, Financial Incentives and Health." *Labour Economics* 6(2): 203–27.

Koning, Pierre. 2009. "Experience Rating and the Inflow into Disability Insurance." *De Economist* 157(3): 315–35.

Koning, Pierre, and Daniel van Vuuren. 2007. "Hidden Unemployment in Disability Insurance." *Labour* 21(4): 611–36.

Koning, Pierre, and Daniel van Vuuren. 2010. "Disability Insurance and Unemployment Insurance as Substitute Pathways." *Applied Economics* 42(5): 575–88.

Gezondheidsenquête. 2010. Health Survey of 2010. Statistics Netherlands.

OECD. 2010. *Sickness, Disability, and Work: Breaking the Barriers; A Synthesis of Findings across OECD Countries*. Organisation for Economic Co-operation and Development.

OECD. 2011. *Sick on the Job. Myths and Realities about Mental Health and Work*. Organisation for Economic Co-operation and Development.

Parsons, Donald O. 1991. "Self-Screening in Targeted Public Transfer Programs." *Journal of Political Economy* 99(4): 859–76.

Permanente Onderzoek Leefsituatie (POLS). Data from 1998, 2002, and 2006. Statistics Netherlands.

Sonsbeek, Jan-Maarten van, and Raymond H. J. M. Gradus. 2013. "Estimating the Effects of Recent Disability Reforms in the Netherlands." *Oxford Economic Papers* 65(4): 832–55.

Tompa, Emile, Kim Cullen, and Chris McLeod. 2012. "Update on a Systematic Literature Review on the Effectiveness of Experience Rating." *Policy and Practice in Health and Safety*, issue 2, pp. 47–65.

UWV. 2007. "UWV Kwartaal Verkenning 2007-III". Amsterdam: Uitvoeringsinstituut Werknemersverzekeringenb (UWV). http://www.uwv.nl/overuwv/Images/UKV%202007-III%20(4%20okt%202007).pdf.

UWV. 2012. *Statistische Tijdreeksen UWV 2012.* Amsterdam: Uitvoeringsinstituut Werknemersverzekeringenb. http://www.uwv.nl/overuwv/Images/Statistische_tijdreeksen_2012.pdf.

UWV. 2013. "Kwantitatieve informatie 2012". Amsterdam: Uitvoeringsinstituut Werknemersverzekeringenb (UWV). http://www.uwv.nl/overuwv/Images/Kwantitatieve%2informatie%202012.pdf.

Vuren, Annemiek van, and Daniel J. van Vuuren. 2007. "Financial Incentives in Disability Insurance in the Netherlands." *De Economist* 155(1): 73–98.

Journal of Economic Perspectives—Volume 29, Number 2—Spring 2015—Pages 173–190

Disability Benefit Receipt and Reform: Reconciling Trends in the United Kingdom[†]

James Banks, Richard Blundell, and Carl Emmerson

In the United Kingdom, public spending on total disability benefits rose steadily from about 0.4 percent of national income in 1950 to about 0.9 percent of national income in the 1980s. This was a period without significant reforms in the program other than a reform of the benefit rate structure in 1971. After this period, as shown in Figure 1, total spending on disability benefits increased sharply over the first half of the 1990s, reaching 1.6 percent of national income in 1995–1996. Concern with this spending triggered a major reform of the UK disability system that came into effect in that same year. As a direct result, spending on disability benefits fell both in real terms and as a share of national income. Further reforms took place over the 2000s, and UK public spending on disability benefits has continued to fall. Public spending forecasts for these disability benefits in 2018–2019 project them to be at their lowest level as a share of national income since the late 1960s.

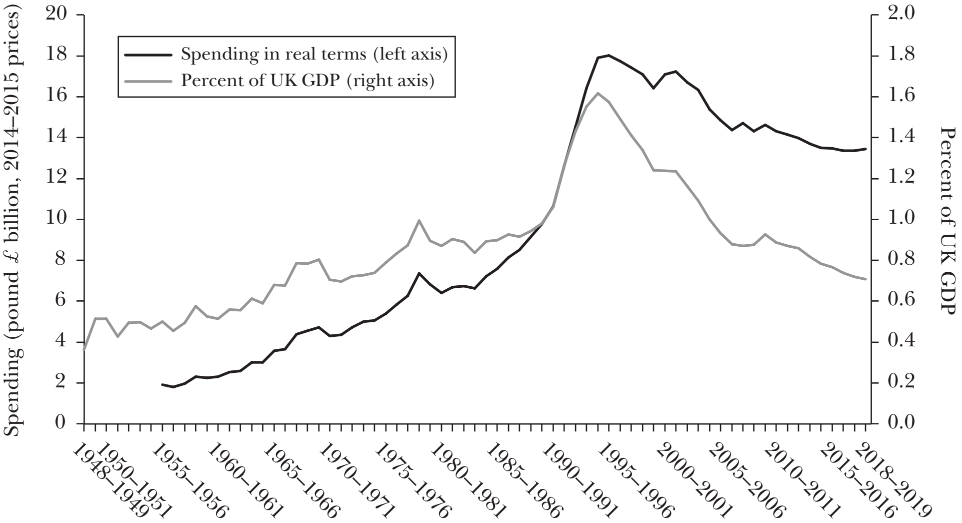
Despite these falls in public spending on disability benefits since the mid 1990s, the numbers in receipt remain high by historical standards. At the end of 2013, 2.3 million individuals in Great Britain were receiving disability benefits, and while this was lower than the 2.5 million recipients of these benefits in 1995, the total

■ *James Banks is Professor of Economics, University of Manchester, Manchester, United Kingdom, and Deputy Research Director of the Institute for Fiscal Studies (IFS), London, United Kingdom. Richard Blundell is the David Ricardo Chair of Political Economy at University College London, and Research Director, Institute for Fiscal Studies, both in London, United Kingdom. Carl Emmerson is Deputy Director, Institute for Fiscal Studies, London, United Kingdom. Their email addresses are j.banks@ifs.org.uk, r.blundell@ucl.ac.uk, and carl_emmerson@ifs.org.uk.*

[†] To access the Data Appendix, visit
http://dx.doi.org/10.1257/jep.29.2.173

doi=10.1257/jep.29.2.173

Figure 1
Total Spending on Disability Benefits in Great Britain, 1948–49 to 2018–19



Source: Department for Work and Pensions, Benefit Expenditure and Caseload Tables, March 2014 (https://www.gov.uk/government/publications/benefit-expenditure-and-caseload-tables-2014).
Note: Figure shows spending on Sickness Benefit, Invalidity Benefit, Severe Disablement Allowance, income support on grounds of disability, Incapacity Benefit, and Employment and Support Allowance.

was still higher than any year prior to the mid 1990s and more than twice the level seen in any year in the 1970s or the first half of the 1980s.[1] In particular, while the period since the mid 1990s has seen a decline in the number of men aged 50 to 64 receiving disability benefits (from 1.1 million in 1996 to 0.6 million in 2013), the number of women aged 16 to 59 receiving these benefits has grown (from 0.8 million to 1.0 million). These trends mean that the previous government's aspiration to reduce the number of disability benefit claimants by one million over the decade starting in 2006 (Department for Work and Pensions 2006) will likely be missed by some considerable distance.

This paper analyzes these and other trends in UK disability recipients and payments trends using administrative data sources alongside an analysis of newly available household survey data and places the trends in the context of the major reforms that have been implemented over the period from 1971 to the present time.[2]

[1] Source: UK Department for Work and Pensions administrative data for data from February 2001 to November 2013 inclusive. Data from 1971 to 1998 are taken from Anyadike-Danes and McVicar (2008).
[2] One challenge for this analysis is the differing geographical coverage of the data that are available to us. Administrative data from the Department for Work and Pensions relate to Great Britain (that is, not including Northern Ireland), while the Labour Force Survey (LFS) data cover the whole of the United Kingdom. English Longitudinal Study of Ageing (ELSA) data, on the other hand, cover England, which has, on average, slightly lower rates of disability benefit receipt than the rest of the United Kingdom.

The analysis in this paper is descriptive in nature, and some of the earlier trends are documented in previous studies (for example, Disney and Webb 1991; Anyadike-Danes and McVicar 2008). However, this paper makes a contribution in the following three ways.

First, it brings together all available data over the longest possible time period into a single set of evidence focused on the key trends in disability benefit receipt. For example, by combining data on age, education, and benefit receipt we can see that, by the end of our sample period in 2013, low-education 25–34 year-olds were twice as likely to be on disability benefits as the highest-education 55–64 year-olds.

Second, to our knowledge, this paper is the first to examine recent trends in receipt of disability benefits by health status; in particular we use administrative data that are available back to 1999 and survey data on those aged 50 and over containing an objective measure of health and disability from 2002 onwards. For example, among both men and women there is systematic growth in the proportion of claimants in any age group with mental and behavioral disorders as their principal health condition, posing an increasingly central issue for future disability policy reform.

Third, this paper is the first, to our knowledge, that documents the recent trends in receipt of disability benefits over the period in which the most recent major UK reform is being implemented.

It seems clear that in the absence of the various reforms discussed here, the number receiving disability benefits in the United Kingdom—and the amount spent publicly on them—would have ended up being substantially higher. But the changes in receipt of disability benefits are far from uniform across divisions of age, sex, education, and health.

The UK Disability Reforms

The main UK disability payment program has changed its name over time. In 1971, it was referred to as *Invalidity Benefit*. The reforms of 1995 changed this to *Incapacity Benefit*. In 2006, the "pathways-to-work scheme" was introduced, which led to the replacement of Incapacity Benefit with the *Employment and Support Allowance* in 2008. This section provides some key details of these benefits and the relevant reforms; for a more detailed discussion, see Banks, Blundell, Bozio, and Emmerson (2012).[3] These benefits are intended for those whose health means that they are not (currently) able to carry out paid work. While individuals do need to have made a contribution to disability insurance through a payroll tax in order to be eligible for benefits, the link from the amount paid to the disability benefits received is

[3] Other substantial reforms to disability benefits in the United Kingdom include: the introduction of Invalidity Benefit in 1971; the introduction of statutory sick pay in 1983; a tightening of the contribution requirements, an intended tightening of the health test (and renaming it the "Personal Capability Assessment," PCA), and means-testing receipt of contributory Incapacity Benefit against an individual's private pension income in 2000; and the time-limiting of receipt of contributory Employment and Support Allowance for those in the work-related activities group to one year in 2012.

weak: generally speaking, the disability payments are flat-rate (regardless of the level of previous earnings), and those with low family incomes can qualify for a means-tested benefit of the same value without having made any earlier contribution.

The value of UK disability benefits relative to earnings peaked in the late 1970s at around 25 percent of average earnings (across all workers), which is a lower level of earnings replacement than provided by disability benefits in the Netherlands, Spain, or the United States (Wise 2014). Since the early 1980s, UK disability benefits have been raised in line with prices; this led the value of these benefits to fall to around 15 percent of average earnings (Banks, Emmerson, and Tetlow 2014) as average earnings in the UK, at least until recently, have tended to rise in real terms over time. As a result, those with average or higher earnings in the United Kingdom receive little protection from publicly provided disability insurance because the amount of disability benefit they could receive is significantly lower than their current earnings.

When Invalidity Benefit was replaced with Incapacity Benefit (in 1995), the main difference was that the "suitable work test" (applied after 28 weeks of incapacity) was replaced with an "all work test." This change meant that, for example, an individual who had been working as an economics professor would be assessed according to the ability of that person to do any kind of paid work rather than the ability to do work that might be considered appropriate for that person's skills and experience. In addition, the new medical screening was removed from the personal doctors of the workers and was administered instead by medical staff at the regional level commissioned by the scheme's administration. Finally, whereas previously those reaching the State Pension age—until April 2010, this was age 65 for men and age 60 for women—could choose to continue receiving Invalidity Benefit for up to five more years instead of moving onto their State Pension, in 1995 this option was removed.

The effect of not allowing those reaching the State Pension age to continue receiving disability benefits for up to five years (instead of the State Pension) led to the number of men aged 65 to 69 and women aged 60 to 64 receiving disability benefits to fall to (effectively) zero over the five years from 1995. It would be easy to overstate the reduction in public expenditure brought about by this reform because the vast majority of these individuals would have instead received the State Pension, which would be payable at a similar amount for many.[4]

While the 1995 reform had, for working-age claimants, been focused on reducing the flow onto the disability benefit, a pair of reforms in the 2000s had the additional aim of increasing the flow off disability benefits. The "pathways-to-work" pilot programs, which began in October 2003, compelled disability benefit recipients to attend a set of interviews focused on how they might better manage their health condition in order to be able to return to paid work and provided them

[4] A similar interaction effect in reverse is seen as the female State Pension age rose from 60 in April 2010 to 62 by May 2014, with 88,000 women aged 60 or over in receipt of disability benefits by the latter date. However, since the increasing State Pension age has affected all women in the cohort and only a subset have moved onto disability benefits, the rise in the State Pension age has led to a overall reduction in spending on State Pensions greater than the increase in disability benefit payments and hence a net strengthening of the public finances (Cribb, Emmerson, and Tetlow 2013).

with an additional (time-limited) financial incentive to move off disability benefits and into paid work. These pilot programs were aimed initially at those who had just moved onto the benefit. The evaluation evidence suggested that the reform was effective at moving individuals off these benefits and also effective at moving them back into paid work (Adam, Bozio, Emmerson, Greenberg, and Knight 2008). The reforms then were extended nationwide and also applied to existing claimants, although the evidence on the effectiveness of these extensions is more mixed (Bewley, Dorsett, and Ratto 2008; Bewley, Dorsett, and Sallis 2009).

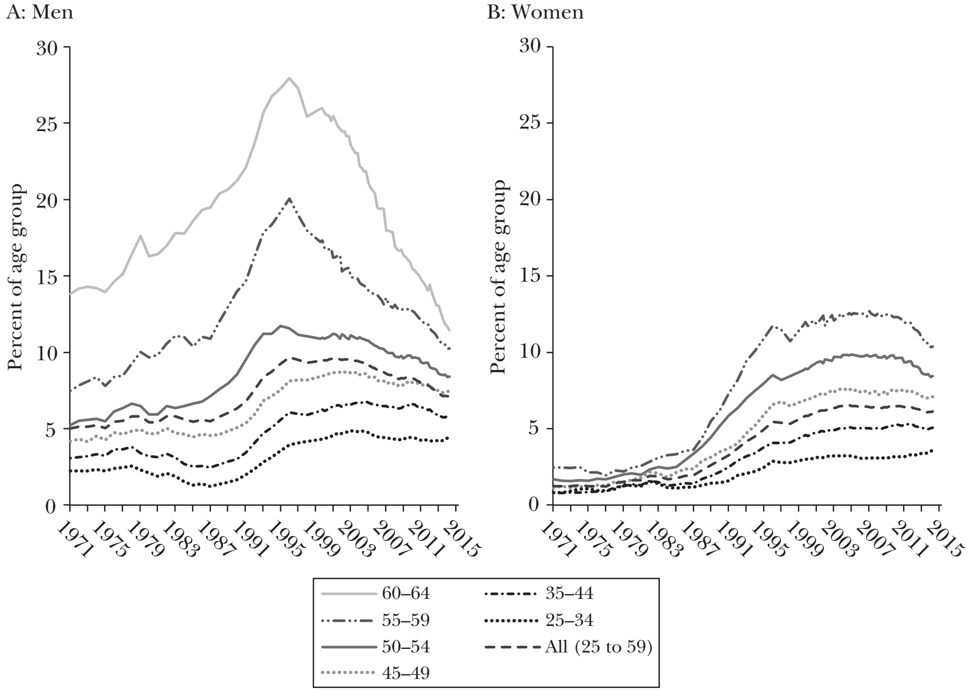
From October 2008, Incapacity Benefit for new claimants was replaced by the Employment and Support Allowance (ESA). The health test for Employment and Support Allowance is intended to be stricter than the health test for Incapacity Benefit: specifically, the new "Work Capability Assessment" (WCA) splits successful claimants for disability benefit into those deemed to have "limited capacity to work and no ability to follow work-related activities" and the remainder who have "limited capacity to work but are able to follow work-related activities." Claimants in the latter group are required to attend the (now nationwide) pathways-to-work program and attend regular interviews with an advisor to discuss, for example, job goals and skill enhancement. In October 2010, reassessment of the stock of existing claimants of Incapacity Benefit began on a trial basis in order to move some to Employment and Support Allowance; the process was then was rolled out nationwide starting in April 2011 (beginning with those with the shortest Incapacity Benefit durations and moving through to the longest) and was to be completed in 2014.

Disability Patterns over Time by Age, Sex, and Reason: Evidence from Administrative Data

Against the background of these reforms in the last few decades, how have the patterns of those receiving disability benefits been evolving? As shown in Figure 2, the rate of disability benefit receipt is greater among older age groups than younger age groups, which is of course entirely unsurprising. Among men aged 50 and over, rates of disability benefit receipt peaked in the mid 1990s, with receipt among those aged 55 to 59 falling from 20 percent in the mid 1990s to just over 10 percent now, while receipt among those aged 60 to 64 fell even more sharply over the same period. While receipt of disability benefits among men aged 25 to 44 continued to rise after the 1995 reform, it rose less quickly than it had been prior to 1995. Similarly, growth in rates of receipt of disability benefits among all age groups of women also clearly slowed after the 1995 reform.

The effect of the replacement of Incapacity Benefit with Employment and Support Allowance on rates of disability benefit receipt appears less clear. Receipt of disability benefits among men aged 35 to 49, and among women aged 50 to 59, appears to have begun falling during the period that Employment and Support Allowance has been rolled out nationwide (2011–2014), whereas prior to this, receipt among these groups had been either flat or rising over time. One might

Figure 2
Disability Benefit Claimant Rates of Men Aged 25 to 64 by Age Group, 1971 to 2014



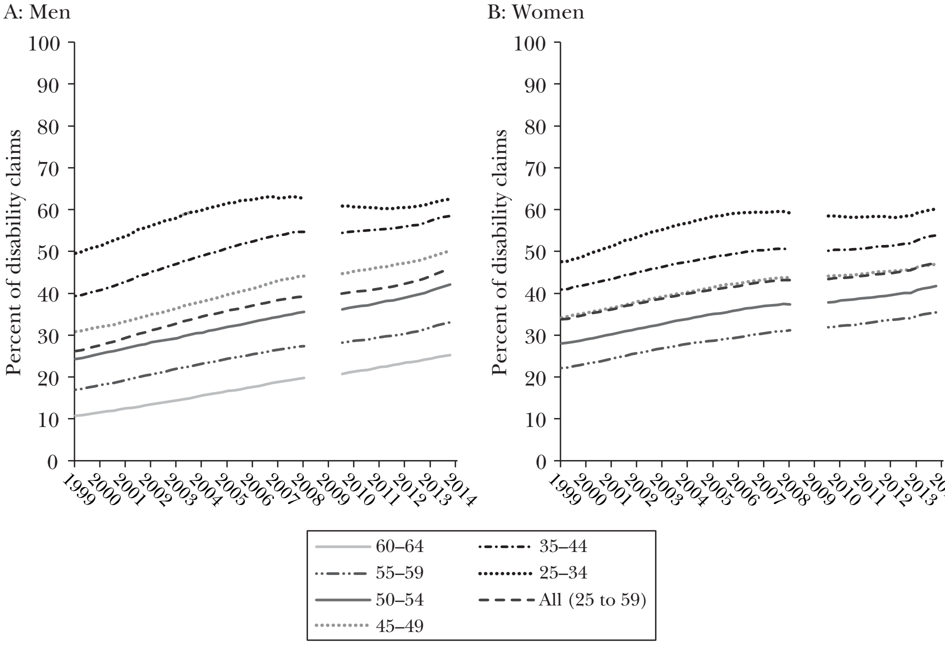
Sources: Authors' calculations using data from the Department for Work and Pensions tabulation tool (http://tabulation-tool.dwp.gov.uk/100pc/tabtool.html, accessed on November 14, 2014), for data from August 1999 to May 2014 (inclusive). Data from 1971 to 1998 are taken from Anyadike-Danes and McVicar (2008). Population estimates (to 2012) and projections (for 2013) by age are taken from the Office for National Statistics.
Notes: Claimants of Invalidity Benefit, Incapacity Benefit, and Employment and Support Allowance in Great Britain. No data are shown for women aged 60 to 64 since age 60 was the State Pension age for women for most of this period. For comparability the "All" category for both men and women is shown for those aged 25 to 59.

have expected that these claim rates would increase during the recession, so the evidence suggests that the most recent reforms have reduced the share of those receiving disability benefits—albeit to a much lesser extent than the 1995 reform.

While the rate of receipt of disability benefits has been declining sharply for older men since 1995 and older women since 2008, the rate of receipt among these groups still remains higher than among other groups. And when combined with the increasing size of this "baby boom" cohort over the last 20 years, this means that there have not been large declines in the absolute numbers receiving disability benefits. Administrative data show the number of disability claimants as 2.52 million in 1996, falling only to 2.29 million by 2013. But when looking at the probabilities of receipt by age, the patterns in Figure 2 suggest that the number receiving disability benefits—and correspondingly the amount of public funding spent on

Figure 3

Percent of Disability Benefit Claims due to Mental and Behavioral Disorders, by Sex and Age Group, August 1999 to May 2014



Source: Authors' calculations using data from the Department for Work and Pensions tabulation tool (http://tabulation-tool.dwp.gov.uk/100pc/tabtool.html, accessed on November 21, 2014).
Notes: Claimants of Incapacity Benefit and Employment and Support Allowance in Great Britain. Data from November 2008 to November 2009 (inclusive) are missing as the summary disease code for those receiving Employment and Support Allowance are not available over this period. The figure shows men aged 25 to 64 and women aged 25 to 59 as ages 65 and 60 were the State Pension ages for men and women respectively for most of this period.

them—would have ended up being substantially higher in the absence of the 1995 and (to a lesser extent) subsequent reforms.[5]

The administrative data also provide evidence on the principal health condition that has led to the disability benefit claim, with data available from August 1999 onwards. The share of disability benefit claims for reasons relating to mental or behavioral health problems, by age group and sex, are shown in Figure 3. Because physical health problems become relatively more prevalent at older working ages, a higher proportion of the disability claims of younger men and women are for reasons relating to mental and behavioral disorders than is the case among older men and women.

[5] In addition, the fact that the 1995 reform essentially disqualified men and women above the State Pension age from receiving benefits means it would have had an additional effect on the number of benefits even though this may not have been associated with reduced government spending for the reasons described above relating to substitution with State Pension expenditures.

The striking pattern is that, for all age and sex groups, a steadily increasing proportion of disability benefit claims are primarily for mental and behavioral reasons. For example, among men aged 50 to 54, this proportion increased from 24 percent in August 1999 to 42 percent in May 2014, while among women of the same age group, this increased from 28 to 42 percent over the same period. If it seems unlikely that mental and behavioral disorders have become this much more prevalent over a relatively short time period, then it would follow that individuals are becoming more likely to be deemed eligible for disability benefits on these grounds. Also, this trend may suggest consequences for which different employer or occupational health adaptations are required to facilitate these individuals returning to paid work.

Two other patterns in Figure 3 merit some comment. First, over the period 2010–2012 the proportion of disability benefit claims of men and women aged 25 to 34 that are primarily for mental and behavioral reasons stopped increasing. While this coincided with the national rollout of the Employment and Support Allowance to the stock of previous Invalidity Benefit claimants with the shortest benefit durations, growth in the proportion of recipients in this age range claiming for mental and behavioral reasons subsequently returned in 2013. Second, the gradient across age has narrowed slightly over time for men, bringing it towards that observed for women, perhaps suggesting that the issues involved with getting disability benefit recipients back into paid work may not now be that different between men and women.

Education, Health, and Disability: Evidence from Survey Data

The receipt of disability benefits and the changes in receipt over time do not just vary by age and sex. There are also important differences across subgroups defined by education and health status. Given the low rates of the benefit levels (now less than 15 percent of average earnings), earnings replacement rates are substantially lower for those with high education than for those with low education and low earnings capacity. Thus, one might expect different reactions to the benefit reforms from different groups. Similarly, the correlations between health, disability, education, and wages make it important that we investigate trends by health and education jointly.

Administrative data are not sufficiently detailed to document such differences, so in this section, we turn to evidence from two household surveys: the Labour Force Survey (LFS) for data on disability by education and the English Longitudinal Study of Ageing (ELSA) for data on disability by health and education status jointly.[6]

[6] Prior to these dates, the data available are more limited either in terms of sample sizes within year-age-education cells or in terms of the nature and detail of the information collected on either benefit receipt or on health. However, some of these other data sources like the British Household Panel Study, the Family Expenditure Survey, and the Health Survey for England have been used to examine specific issues and questions pertaining to disability benefits in earlier years. For example, Disney, Emmerson, and Wakefield (2006) use the BHPS to examine the importance of health in determining labor market transitions of working-age individuals aged 50 and over.

The LFS covers the full UK household population and samples an average of around 250,000 individuals between ages 25–59 for women and ages 25–64 for men per year over the period 1998 to 2013. These data allow us to examine the extent to which, since 1998 at least, these trends in receipt of disability benefits have differed across low- and high-skilled individuals within each age group. In drawing this sample, we excluded individuals aged below 25 since a nonnegligible and endogenous fraction will still be in full-time education. We also excluded those at or above the State Pension age or one year below it: that is, women aged 59 and over and men aged 64 and over since they were not eligible for benefits over this sample period.

We define three education groups. The lowest-education group are those that left full-time education at or before the compulsory school-leaving age (age 14 for those born before April 1933, age 15 for those born between April 1933 and April 1958, and age 16 otherwise). This represents a large fraction of the population, particularly in older cohorts.[7] The remaining individuals are divided into a middle-education group who chose some post-compulsory schooling but left school at or before age 19 and a high-education group who continued full-time education beyond age 19. The LFS data also contain standard and relatively high-quality measures of disability benefit receipt status.

Those with lower levels of education are more likely to claim disability, as one would expect, for several reasons. There is a well-documented relationship between lower levels of education and worse health. In the UK system, there is also the fact that the flat-rate structure of benefit rates, described above, means that disability benefits will replace a higher share of earnings for lower earning groups—thus making disability a substantially less attractive option for those with medium or higher education levels. These differences are immediately apparent in the LFS data: within all age and sex subgroups, the low educated are roughly four times more likely to be in receipt of disability benefits than their high-education counterparts.

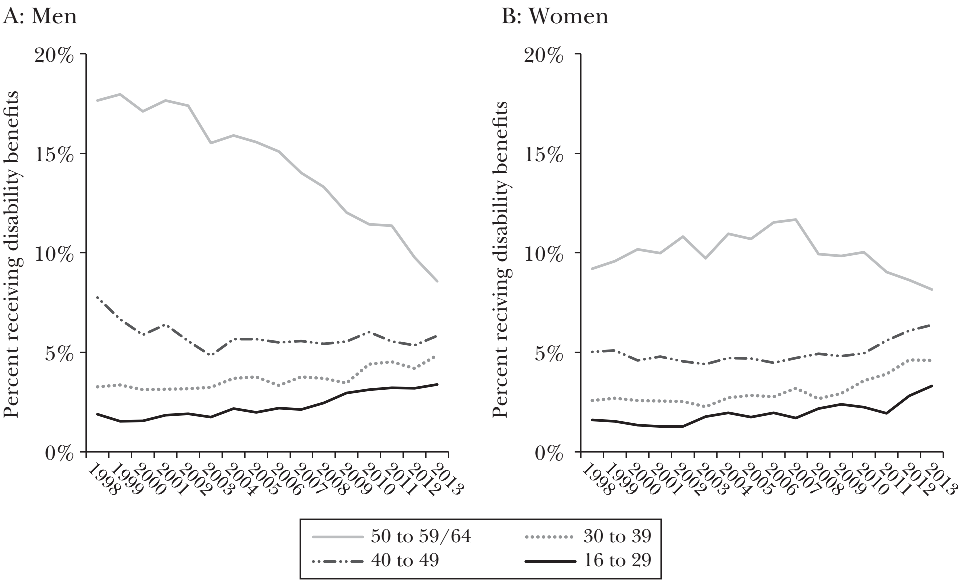
In addition, patterns over the last few years in particular need to be interpreted with caution because the recent recession is known to have affected education and sex subgroups differently. For example, the most recent recession involved a smaller fall in the proportion of paid work than in previous UK recessions, and this difference was particularly striking for older, better-educated male workers (Blundell, Crawford, and Jin 2014). With that caution duly noted, some patterns of disability across education and age levels still stand out.

While the reductions in rates of disability benefit receipt since the 1995 reform have come mainly from older men (as noted earlier), amongst this group the trend has been considerably more acute for the lower educated than for the higher educated. And across older men and women, there appears to have been a differential trend in disability levels across education groups over the period 2010–2013, particularly for older women, whereby the probability of receipt has fallen particularly rapidly for the lowest-educated older women and yet been relatively flat

[7] The cohort with school-leaving age of 14 are not actually observed below State Pension age from 1998 onwards so do not feature in our analysis.

Figure 4

Receipt of Disability Benefits among Those with a Low Level of Education, by Sex and Age Band, 1998 to 2013



Source: Authors' calculations using data from the Labour Force Survey.
Notes: Claimants of Incapacity Benefit and Employment and Support Allowance. Figure shows men up to age 64 and women up to age 59 since ages 65 and 60 were the State Pension ages for men and women respectively for most of this period.

for more-educated older groups and either flat or rising for the more-educated younger groups. Figure 4 shows the stark differences in patterns for low-education older groups in comparison to their younger counterparts. Furthermore, while we do not present a figure on this, differential levels of disability across education levels have been widening in the last few years for those in the 25 to 44 year-old age brackets, particularly for women in the period post-2008 from when the Employment and Support Allowance began to be introduced.

The combination of these patterns of disability across age and education groups over time, and in particular the sharp declines for the oldest age groups following the recent reforms, mean that for both men and women by the end of our sample period in 2013, low-education 25–34-year-olds are in fact more than twice as likely to be on disability benefits than the highest-education 55–59/64 year-olds.[8] This

[8] Running a simple logistic model of benefit receipt on age, sex, and education by year, we find that, in 2013, those with high levels of education aged 50 to 59/64 are found to be half as likely to be in receipt of disability benefits as those with low levels of education aged 25 to 34 (odds ratio 0.49 with a 95 percent confidence interval of 0.41 to 0.59), whereas in 1998 they were twice as likely (odds ratio 2.15 with a 95 percent confidence interval of 1.81 to 2.56).

pattern may be related to the fact, noted earlier, that for those with low-earnings potential, the relative payoff to work versus disability benefits will be lower, with this effect growing over time as benefits are updated with the price level (Banks, Emmerson, and Tetlow 2014).

Because the differences in trends across education within the oldest age groups have been so stark, it is important to understand these in more detail, and in particular, their relationship with health and employment rates. For this purpose, we turn to data from the English Longitudinal Study of Ageing (ELSA), although this means that we can only focus on the older age groups. The first six waves of ELSA have been collecting highly detailed information on the health, functioning, and socioeconomic status of around 11,000 individuals aged 50 and over in England since 2002. Individuals are interviewed every two years with an additional nurse assessment every four years. The resulting data encompass self-reported measures of disability, physical, and cognitive performance tests, and many other detailed biomarker and self-reported health indicators, as well as the standardized questions on education and self-reported receipt of disability benefits we utilized from the LFS data.

The ELSA data would, of course, support a highly-detailed modeling of the dynamics of disability and disability benefit entitlement, especially because of the longitudinal nature of the panel data and because many individuals have given consent for their data to be linked to their administrative benefit records. Here, however, we confine ourselves to constructing a simple index of health and disability based on those disability conditions that are covered by the ELSA data and also assessed as part of the Work Capacity Assessment. The measures relate to physical disability (and, thus, to walking, standing, sitting, manual dexterity, and so on), vision, incontinence, mental health, and finally stress-related reasons for leaving past employment. In total, there are 11 such items, so we construct an index, taking values 0–11, which simply counts the number of conditions each individual in the ELSA sample is observed to have.[9] Having looked at the distribution of the index, and in order to keep our descriptive analysis as simple as possible, we then group the data into four disability categories: None (0 or 1 condition), mild (2 or 3 conditions), moderate (4 or 5 conditions), or severe (6 or more). Other than for reasons of sample size, our subsequent conclusions are not affected materially by using different groupings. The majority of individuals aged 50 to the State Pension age have zero or one of the identified conditions, although the majority of low-education individuals have one or more. The distribution of the disability index and its covariation with education is strikingly similar in both 2002 and 2012, although the data do show a slight reduction in the proportion of women with the highest values of the disability index over this time period.

[9] More specifically, these conditions relate to various Activities of Daily Living (walking, sitting, standing, climbing stairs, lifting a weight, picking up a 5p coin, etc.), as well as eyesight, incontinence, and stress. Further details are available from the authors on request.

Table 1
Disability Benefit Receipt Rates, by Age, Sex, and Disability Level

Year	*All aged 50–State Pension Age*	*Male, low education*	*Male, high education*	*Female low education*	*Female, high education*
2002					
None (0–1)	2.9	5.5	2.7	1.7	1.4
Mild (2, 3)	15.0	25.7	13.9	11.1	9.8
Moderate (4, 5)	34.4	50.9	33.3	29.8	16.2
Severe (6+)	55.3	72.2	54.3	37.9	45.2
2012					
None (0–1)	0.9	3.0	0.9	0.6	0.1
Mild (2, 3)	6.8	11.2	7.0	5.7	4.4
Moderate (4, 5)	15.5	21.1	18.6	8.2	14.7
Severe (6+)	37.2	51.8	34.2	31.1	31.7

Notes: Authors' calculations from waves 1 and 6 of the English Longitudinal Study of Ageing. 2002 numbers are receipt of Invalidity Benefit; 2012 numbers are receipt of either Invalidity Benefit or Employment and Support Allowance.

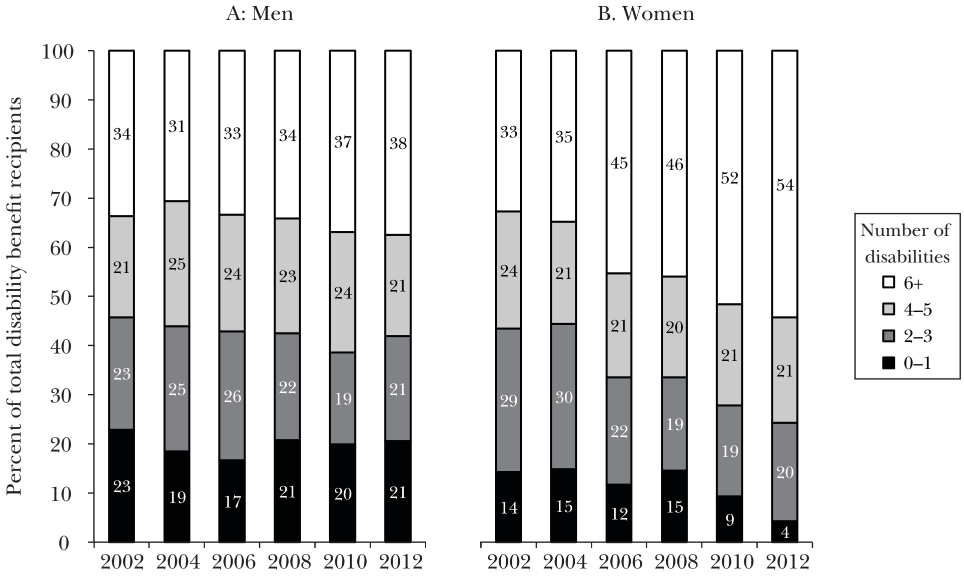
While this index is crude, it still correlates strongly with the receipt of disability benefits particularly for men: within the age group between 50 and the State Pension age, almost three-quarters of men with low education and severe disability (six or more conditions) in 2002 are observed to be receiving disability benefits. For a given level of the disability index, the probability of disability benefit receipt is greater for men than for women (perhaps unsurprising giving lower lifetime labor market attachment for women in these cohorts) and greater for those with low levels of education than for those with high levels of education. Again, this may be related to the relatively low replacement rate of disability benefits for higher earners.

However, the probability of receipt of disability benefits conditional on health status did change substantially between 2002 and 2012, as is also apparent in Table 1. Declines in benefit receipt are observed at all values of the index, with large absolute declines in receipt rates in the groups with the worst level of disability, and large proportionate declines in receipt rates for the least disabled. These trends are consistent with an improved targeting of the benefit onto those with more severe disabilities, although this improved targeting happened at different points of the disability distribution before and after the 2008 reform when the Employment and Support Allowance first began to replace Incapacity Benefit.

Figure 5 shows how the distribution of disability within the stock of benefit recipients changed over the period 2002–2012, as a result of these changing rates of receipt conditional on health and the changing distribution of the health index more generally. All groups experienced a reduction in the proportion receiving disability benefits, but these reductions were much greater for some groups than for others.

Figure 5

Composition of Disability Benefit Recipients by Sex and Disability Index, Individuals Aged 50 to State Pension Age, 2002 to 2012



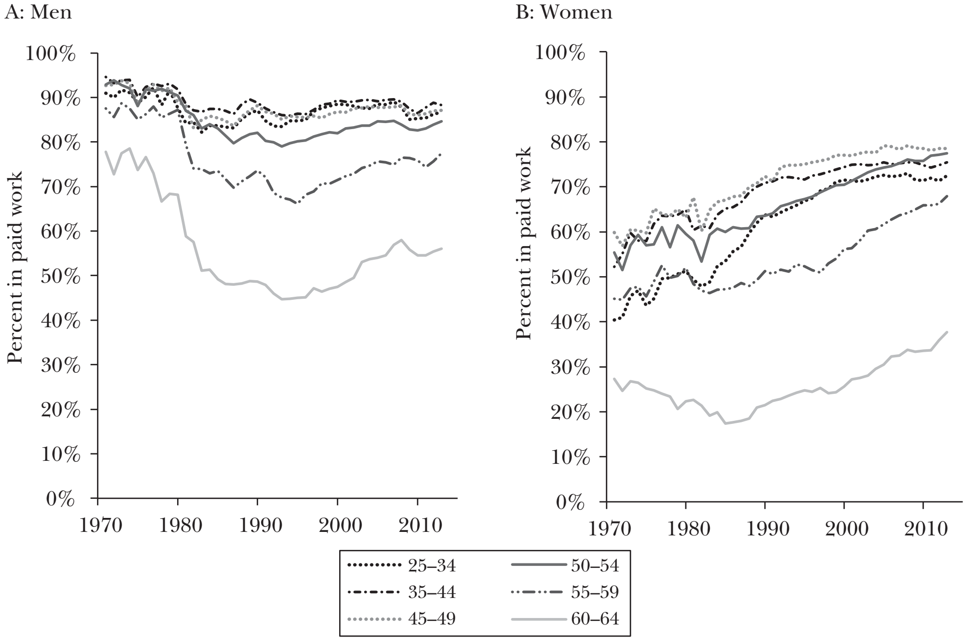
Source: Authors' calculations from waves 1 and 6 of the English Longitudinal Study of Ageing.
Notes: An index of 0–1 means 0 to 1 disability conditions, 2–3 means 2 or 3 conditions, and so forth. See text for details.

Among women, the proportion of benefit recipients with six or more indicated disabilities increased over the whole period from one-third to over one-half of the total. In the most recent years, subsequent to the Employment and Support Allowance reform, there is striking reduction in the proportion of benefits with zero or one conditions, falling from 15 to 4 percent over the four-year period 2008–2012. For men, the disability benefit appears less well targeted at the beginning of the period: 23 percent of male benefit recipients in 2002 reported zero or one condition in the health index compared to only 14 percent of women. After that time, there are no marked or consistent trends in the disability composition of the male benefit claimants. Thus, it seems that the substantial reduction in rates of disability receipt for older individuals documented in the previous analysis has been less driven by the best-health groups for men than it has been for women.

Finally, we compare trends in employment with the disability trends both unconditionally and by education and health groups, to offer some tentative evidence relating to the question of substitution between disability benefit and employment. The rules for receiving disability benefits offer only very limited possibilities for doing any paid work at all (certain permitted work rules do allow benefit recipients to do a small amount of paid work each week while receiving benefits only if their

Figure 6

Percent Aged 25 to 64 in Paid Work by Sex and Broad Age Group, 1971 to 2013



Source: Authors' calculations from pooled Family Expenditure Survey (1968–2012, N = 390,477) and Labour Force Survey (1975, 1977, 1979, 1983–2013, N = 6,997,526) microdata. Estimates for years with only FES data are adjusted based on age-sex specific relationship between FES and LFS rates in the years up to and including 1983 when both surveys were in place.

Notes: Definition of work includes all those working in the reference week (LFS) or month (FES) either as full or part-time employees or self-employed. Excludes babysitting coded as self-employment in FES pre-1982.

Jobcentre Plus adviser agrees).[10] Thus, any trends toward lower claiming of disability benefits that are not matched by trends in higher employment suggest either movements onto other welfare programs or else the use of other forms of support (like family income or savings) until the individual reaches the State Pension age. This outcome would also indicate a broad failure of the reforms if assessed with respect to the goal of moving people into paid work, as opposed to simply reducing caseloads or government spending on disability benefit payments.

The broad trends in labor force participation in the United Kingdom presented in Figure 6 are well known and similar to those in the United States. For example, there has been a steady increase in paid work for men of older working ages from

[10] In the English Longitudinal Study of Ageing (ELSA) data, only a small fraction of those on benefits report having done any paid work in the previous month—around 0.3 percent of all males and 0.2 percent of all females over the period 2002–2013. These levels are small enough that they could largely be accounted for by measurement error in the survey.

Table 2
Changes in Benefit Receipt and Employment, by Health Level, 2008–2012

	Percentage point difference (2008–2012) in:				
Disability level	*Disability Benefit receipt*	*Employment*	*Job Support Allowance*	*N 2008*	*N 2012*
Mild (2, 3)	−3.11 (−1.91)	−0.71 (−0.26)	1.04 (1.23)	683	559
Moderate (4, 5)	−17.83 (−4.64)	3.23 (0.74)	1.62 (1.30)	234	245
Severe (6+)	−20.62 (−4.49)	−0.98 (−0.31)	−0.20 (0.14)	237	215

Notes: Authors' calculations from waves 4 and 6 of the English Longitudinal Study of Ageing. Table presents difference in cell means between 2008 and 2012, with all variables adjusted for the sex- and education-specific trends in the group with 0–1 health and disability problems. Numbers in parentheses are *t*-ratios

the mid 1990s onwards, and steadily rising rates for women of all ages. At the aggregate level, it seems that there are at least some examples of broad correspondence in the patterns with receipt of disability. For example, the pick-up in male employment rates has been most apparent for the oldest ages where disability benefit rates have declined the most. In labor market trends for women, there is evidence of an upturn in the trend in employment for 55–59 year-old women since 2010 (as the most recent reform was being implemented) at the same time as the sharp downturn in their benefit receipt rates. But separating the effects of general labor market trends (and, in the case of women, the particular experiences associated with the most recent post-recession period) from any effects due to disability or disability benefits is not straightforward, at least in the most recent years of data, and estimating a causal effect of benefit reform on employment rates is left for future work.

As a final piece of descriptive evidence, however, we disaggregate these trends in receipt of disability further within the population 50 and over—specifically by education, sex, and disability level—and focus just on the years since the Employment and Support Allowance first began to be introduced in 2008. We then examine whether the groups with greater or lesser movement off disability benefits since this reform has occurred have seen greater or lesser movements into work. Table 2 shows the difference in benefit receipt rates between 2008 and 2012 for each of the three levels of disability, along with corresponding changes in employment and in receipt of Job Seekers Allowance (that is, an unemployment benefit), with these differences being expressed relative to changes in the group with zero or one disability condition.[11] Thus, the group with a mild level of disability experienced, on average, a 3 percentage point relative decline in those receiving disability benefits.

[11] More specifically, for each variable in the table we express each observation relative to the mean of those with the same sex and education but without any disability.

At the same time, there was a small (0.7 percentage point) but insignificant decline in employment and only a 1 percentage point rise in those receiving unemployment benefits (also statistically insignificant).

Taken together, the evidence in Table 2 does not provide strong evidence showing effects of the Employment and Support Allowance reform either on return to work or on movements onto unemployment benefits. What evidence there is of substitution between these three forms of activity is strongest for the group with moderate but not severe disability, where we observe higher rates of employment and receipt of Job Seekers Allowance that can account for just under one-quarter of the reduction in numbers on disability benefit. For those with the severest disabilities, we observe a very large relative decline in the numbers on disability benefits (consistent with Table 1) but essentially no change in the numbers either employed or receiving unemployment benefits.

Summary and Conclusions

The United Kingdom has had a number of reforms to the rates and structure of disability benefits in recent decades that make it an interesting case study for other countries thinking of reforming their systems of support for those whose health means they are unable to work. In this paper, we have documented recent reforms and examined the evidence on trends in the numbers and characteristics of those receiving such benefits over the period since 1971. What lessons can we draw from this exercise?

First, the UK experience demonstrates that, at least in time periods after disability benefits have spiked upward, reforms concerning eligibility for such benefits can reduce disability benefit levels from the levels they otherwise would have reached. Given large demographic shifts in cohort size, this is not always apparent when considering raw numbers of claimants.

Second, when calculating potential cost savings from disability reforms, it is important not to consider a single program in isolation. In the United Kingdom, most of the reduction in spending on disability benefits among those over the State Pension age that resulted from the 1995 reform, also resulted in higher State Pension spending. Similar interaction effects will be important when considering how means-tested benefits will rise if the number of working-age adults receiving disability is reduced.

Third, perhaps as a result of the low and declining levels of UK disability benefit in monetary terms, receipt of disability has now become even more closely related to education level than in the past. For example, by the end of our sample period in 2013, low-education 25–34 year-olds are now twice as likely to be on disability benefits as the highest-education 55–64 year-olds.

Fourth, as a greater share of women enter the labor force, a greater share are also eligible for and receiving disability benefits. The decline in the number of older working-age men receiving disability benefits has been partially offset by growth in the number of younger women receiving these benefits. But disability patterns by health and education status are not the same for men and women. The substantial reduction in rates of disability receipt for older individuals documented in recent years has been less driven by a decline among the best-health groups for men than it has been for women. For a given level of reported health status, men are more likely to receive disability benefits than women.

Fifth, there is systematic growth over time in the proportion of claimants in any age and sex group with mental and behavioral disorders as their principal health condition, and the inflow and outflow of this group to disability benefits raises prominent issues for future disability (and employment) policy.

Finally, the evidence with regard to reforms that seek to expedite movements back to employment is mixed. While the period of decline in benefit receipt since 1995 has also been one of increasing employment amongst older age groups, it is not easy to disentangle any effects from broader labor market trends given the limited microdata on health, education, and employment over this time and the nature of the reforms that occurred. Looking tentatively at the more recent Employment and Support Allowance reform as it affects older adults and using its differential impact on groups with differing levels of education and disability, we do not find strong evidence of substitution between disability benefit and unemployment benefits over the period of the most recent reforms, nor do we find strong evidence of any return-to-work effects. But our analysis is very limited, both in its scope and its statistical power, given how recently the reform occurred and the fact that we do not have much data in the post-reform period. Indeed, given the way the program has been rolled out, it is somewhat challenging to think of a more concrete evaluation of these potential substitution effects with the data available, at least until we have a larger sample of those observed with an onset of disability and a (potential) movement from work onto the new benefit. Such an analysis is left as a topic for future research. But the descriptive evidence presented here on how the Employment and Support Allowance reform has had differential impacts on benefit receipt for groups defined by age, sex, education, and disability level suggests that such research may be fruitful in enhancing our understanding of the full effects of disability benefit reform and, indeed, in informing other countries with regard to the potential effects of any future reform.

■ *Funding for this research comes from the Economic and Social Research Council through the Centre for Microeconomic Analysis of Public Policy at IFS (RES-544-28-5001). Data from the Labour Force Survey (LFS) and the English Longitudinal Study of Ageing (ELSA) were made available by the UK Data Archive. ELSA was developed by a team of researchers based at the National Centre for Social Research, University College London, and the Institute for Fiscal Studies. The data were collected by the National Centre for Social Research. ELSA is funded by the National Institute of Aging in the United States and a consortium of UK government departments coordinated by the ESRC. The authors are thankful to James Browne, Antoine Bozio, and Gemma Tetlow for useful discussions in previous projects that have informed much of the analysis here and to David Autor, Chang-Tai Hseih, and Timothy Taylor for useful comments on an earlier draft. Responsibility for interpretation of the data, as well as for any errors, is the authors' alone.*

References

Adam, Stuart, Antoine Bozio, Carl Emmerson, David Greenberg, and Genevieve Knight. 2008. *A Cost-Benefit Analysis of Pathways to Work for New and Repeat Incapacity Benefits Claimants.* Department for Work and Pensions Research Report 498, Department for Work and Pensions, London. http://webarchive.nationalarchives.gov.uk/20130314010347/http://research.dwp.gov.uk/asd/asd5/rports2007-2008/rrep498.pdf.

Anyadike-Danes, Michael, and Duncan McVicar. 2008. "Has the Boom in Incapacity Benefit Claimant Numbers Passed Its Peak?" *Fiscal Studies* 29(4): 415–34.

Banks, James, Richard Blundell, Antoine Bozio, and Carl Emmerson. 2012. "Disability, Health and Retirement in the United Kingdom." In *Social Security Programs and Retirement around the World: Historical Trends in Mortality and Health, Employment, and Disability Insurance Participation and Reforms*, edited by David Wise, 41–77. University of Chicago Press.

Banks, James, Carl Emmerson, and Gemma C. Tetlow. 2014. "Effect of Pensions and Disability Benefits on Retirement in the UK." NBER Working Paper 19907.

Bewley, Helen, Richard Dorsett, and Marisa Ratto. 2008. *Evidence on the Effect of Pathways to Work on Existing Claimants.* Department for Work and Pensions Research Report 488, Department for Work and Pensions, London. http://webarchive.nationalarchives.gov.uk/+/http://www.dwp.gov.uk/asd/asd5/rports2007-2008/rrep488.pdf.

Bewley, Helen, Richard Dorsett, and Sergio Sallis. 2009. *The Impact of Pathways to Work on Work, Earnings and Self-Reported Health in the April 2006 Expansion Areas.* Department for Work and Pensions Research Report 601, Department for Work and Pensions, London. http://webarchive.nationalarchives.gov.uk/20130314010347/http://research.dwp.gov.uk/asd/asd5/rports2009-2010/rrep601.pdf.

Blundell, Richard, Claire Crawford, and Wenchao Jin. 2014. "What Can Wages and Employment Tell Us about the UK's Productivity Puzzle?" *Economic Journal* 124 (576): 377–407.

Cribb, Jonathan, Carl Emmerson, and Gemma Tetlow. 2013. *Incentives, Shocks or Signals: Labour Supply Effects of Increasing the Female State Pension Age in the UK.* Working Paper No. W13/03, Institute for Fiscal Studies, London. http://www.ifs.org.uk/publications/6622.

Department for Work and Pensions. 2006. *A New Deal for Welfare: Empowering People to Work.* Cm 6730. Department for Work and Pensions, London. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/272235/6730.pdf.

Disney, Richard, Carl Emmerson, and Matthew Wakefield. 2006. "Ill Health and Retirement in Britain: A Panel Data-Based Analysis." *Journal of Health Economics* 25(4): 621–649.

Disney, Richard, and Steven Webb. 1991. "Why Are There So Many Long Term Sick in Britain?" *Economic Journal* 101(405): 252–62.

Wise, David A., ed. 2014. *Social Security Programs and Retirement around the World: Disability Insurance Programs and Retirement.* Chicago: University of Chicago Press.

Journal of Economic Perspectives—Volume 29, Number 2—Spring 2015—Pages 191–212

Reforming LIBOR and Other Financial Market Benchmarks

Darrell Duffie and Jeremy C. Stein

In the spring of 2008, LIBOR moved from the fine print of interest-rate contracts to the headlines of newspapers (for example, Mollenkamp 2008; Mollenkamp and Whitehouse 2008). LIBOR is the London Interbank Offered Rate: a measure of the interest rate at which large banks can borrow from one another on an unsecured basis. LIBOR is often used as a benchmark rate—meaning that the interest rates that consumers and businesses pay on trillions of dollars in loans adjust up and down contractually based on movements in LIBOR. Investors also rely on the difference between LIBOR and various risk-free interest rates as a gauge of stress in the banking system. Benchmarks such as LIBOR therefore play a central role in modern financial markets.

Thus, the 2008 news reports revealing widespread manipulation of LIBOR threatened the integrity of this benchmark and lowered trust in financial markets. LIBOR is determined each day—or "fixed"—based not on actual transactions between banks but rather on a poll of a group of banks, each of which is asked to make a judgment-based estimate of the rate at which it could borrow. Banks had incentives to announce biased interest rates, for two reasons. First, in times of economic stress, reporting a lower interest rate would signal that the bank is more creditworthy, all else equal. Second, some of the bank's trading positions would be more profitable if LIBOR could be pushed one way or the other, depending on the position taken.

■ *Darrell Duffie is the Dean Witter Distinguished Professor of Finance and Shanahan Family Faculty Fellow, Graduate School of Business, Stanford University, Stanford, California. Duffie chaired the Market Participants Group on Reforming Interest Rate Benchmarks, established by the Financial Stability Board. Jeremy C. Stein is the Moise Y. Safra Professor of Economics, Harvard University, Cambridge, Massachusetts. Stein co-chaired an Official Sector Steering Group on reference rate reform, while serving as a member of the Federal Reserve Board of Governors. Their email addresses are duffie@stanford.edu and jeremy_stein@harvard.edu.*

http://dx.doi.org/10.1257/jep.29.2.191 doi=10.1257/jep.29.2.191

These problems with LIBOR raised more general issues about benchmarks. Along with LIBOR, there are other "IBORs," including EURIBOR, which is the interbank offered rate at which large banks in the European Union lend to each other, and TIBOR, the Tokyo Interbank Offered Rate, at which large Japanese banks lend to each other. In addition, benchmarks for foreign exchange rates and certain commodity prices appear in many contracts.

The two of us recently contributed to a pair of reports commissioned by the Financial Stability Board that recommend how to make benchmark rates such as LIBOR and other interbank offered rates less vulnerable to manipulation.[1] While these reports cover many technical issues, they are based on two overarching principles.

First, benchmarks should be based—to the greatest practical extent—not on judgments submitted by market participants, but on actual transactions. Anchoring benchmarks in transactions is a key recommendation of several previous policy groups (for example, see International Organization of Securities Commissions 2013). But a tough problem confronts a shift to transaction-based IBOR benchmarks. Remember, the "I" in IBOR stands for "interbank." The daily fixing of LIBOR is supposed to be an estimate of the rate at which major banks can borrow *from each other.* However, there are surprisingly few actual loan transactions between banks that could be used to fix most of the IBORs, including those for the 3- and 6-month maturities that are so widely used as benchmark rates. The thinness of the underlying interbank markets has made it difficult to come up with reliable daily fixings that are transactions-based.

The solution proposed in the policy reports of our groups is to fix the IBORs using a much wider set of unsecured bank-borrowing transactions, not just those in the interbank market. This approach would include rates on "wholesale" (that is, large-denomination) certificates of deposit as well as commercial paper issued by banks to a wide range of nonbank investors.

Second, the reform process should strongly encourage heavier use of alternative benchmark reference rates. The original purpose of the IBORs was to measure average bank borrowing costs, which include a spread component for bank credit risk. Particularly with the enormous boom in interest-rate derivatives trading since the 1980s, IBORs have been heavily used in contracts whose purpose is to transfer risk related to fluctuations in general market-wide interest rates. The motives for these "rates trading" applications generally have little to do with the component of the IBORs that reflects the spread between bank credit and a risk-free interest rate. However, it is a self-reinforcing choice by market participants to trade in more liquid high-volume markets, all else equal. In part through an accident of history, this desire to belong to the high-liquidity club has led to a massive agglomeration of trade based on the IBOR benchmarks.

[1] Duffie chaired a Market Participants Group on Reforming Interest Rate Benchmarks (or the Market Participants Group on Reference Rate Reform). Stein co-chaired (along with Martin Wheatley, head of the UK's Financial Conduct Authority) an Official Sector Steering Group on the same topic, while serving as a member of the Federal Reserve Board of Governors. Both groups were established by the Financial Stability Board. These reports are Market Participants Group (2014) and Official Sector Steering Group (2014).

While such an agglomeration effect is beneficial from the standpoint of liquidity, it increases incentives for market manipulation. The deep and liquid IBOR-based derivatives markets can accommodate extremely large derivatives positions. A trader with a sufficiently large position can profit significantly from even tiny distortions in IBOR fixings, on the order of one basis point (that is, 0.01 percent). In 2008, reporting on the LIBOR scandal revealed that manipulators had arranged for dishonest judgment-based reports of bank borrowing rates. With a transactions-based benchmark, a manipulator might attempt to distort actual transactions. Either way, the message is the same: a thin underlying bank borrowing market cannot be a robust foundation for a multi-hundred-trillion dollar derivatives "rates" market, even with substantial improvements to the IBOR-fixing methodology.

Fortunately, many of the interest-rate trading applications currently served by the IBORs could be as well or better served by risk-free or near-risk-free benchmarks that are not tied to banks' costs of funds. In the United States, for example, interest rates based on Treasury bills or other rates that we will discuss later in this paper (such as general collateral repo rates) would be adequate or preferred for many rates-trading applications.

We do not underestimate the difficulty of getting market participants to opt for alternative reference rates so long as IBOR-based markets are so liquid. Precisely because everybody prefers to be in the high-liquidity club, there is a coordination problem. No individual actor may be willing to switch to an alternative benchmark, even if a world in which many switched would be less vulnerable to manipulation and offer investors a menu of reference rates with a better fit for purpose. Hence, there may be an important role for policymakers to guide markets in the desired direction.

The remainder of this paper is organized as follows. We begin with a discussion of the economic role of benchmarks in reducing market frictions. We explain how manipulation occurs in practice, and illustrate how benchmark definitions and fixing methods can mitigate manipulation. We then turn to an overall policy approach for reducing the susceptibility of LIBOR to manipulation, before focusing on the practical problem of how to make an orderly transition to alternative reference rates, without raising undue legal risks.

The Economics of Benchmarks

Why Use Benchmarks?

Financial market participants rely on benchmarks for a range of purposes that are primarily related to reducing asymmetric information regarding the value of the underlying traded financial instrument.

Consider for illustration a forward contract for gold, committing a buyer to pay the difference between the agreed forward price and the spot price of gold at the future contract settlement date. Without recourse to an independently announced gold price benchmark, the counterparties could easily disagree about the net payment due at the time of settlement. Indeed, the two parties have precisely opposite incentives regarding how to measure the spot price of gold. Thus, without a benchmark, they might expend extra effort to settle their contract. They might avoid contracting based on price, and instead use the more costly (but less manipulation-prone) settlement method of physical delivery of gold. Or in light of the anticipated settlement costs, the two parties might just fail to agree on a contract in the first place, thus losing their gain from trade. Even if a benchmark exists, costs may arise to the extent that the benchmark is not reliably measured or can be manipulated. Indeed, there have been recent allegations of manipulation of gold benchmarks (Vaughn 2014). Clearly, if one of the counterparties to a trade also plays a role in the fixing method that determines the announced benchmark price, the incentive to manipulate is especially severe. This moral hazard may lead to lower market participation or even a market breakdown.

Reliable benchmarks also reduce search costs in bilateral over-the-counter markets, where, in the absence of a centralized exchange, benchmarks can improve matching efficiency and increase participation by less-informed agents. For example, with the publication of an interest rate benchmark such as LIBOR, bank customers are better able to judge whether a loan rate is competitive. Without a benchmark, intermediaries can take greater advantage of market opaqueness and of the cost to customers of searching for alternative quotes. Before the advent of LIBOR in the United States, banks commonly quoted variable-rate loans at some spread above a "prime rate," but each bank decided on its own prime rate, and while these rates moved in relatively close tandem across banks, sophisticated borrowers understood the benefit of shopping around.

In this sort of setting, benchmarks offer financial intermediaries a tradeoff: on one side, benchmarks tend to reduce profit margins; on the other side, this disadvantage can potentially be more than offset through increased volume of trade (Duffie, Dworczak, and Zhu 2014). Thus, intermediaries may find it advantageous to introduce a benchmark. Indeed, in 1969 a consortium of London-based banks led by Manufacturers Hanover introduced LIBOR in order to entice international borrowers such as the Shah of Iran to borrow from them (for a brief history, see Ridley and Jones 2012). By 1984, LIBOR became an official benchmark of the British Bankers Association.

A further transparency benefit of benchmarks applies when investors delegate their trading decisions to agents, who may not always make their best efforts to obtain good trade execution on behalf of their clients. Suppose an investor selling euros for dollars is told by her broker, "We obtained an excellent price of $1.3500 for your Euros." Absent a benchmark, the investor could not easily validate the broker's claim and may be suspicious of the potential for dishonest service. However, if there is a nearly simultaneous published benchmark fixing of an exchange rate of one euro for $1.3501, then the broker's claim of good execution is easily verified. Less-informed investors who delegate their trade execution to agents are thus more willing to participate in markets when incentives for good execution are supported by the existence of reliable benchmarks. The recent report of the Financial Stability Board on foreign exchange benchmarks confirms that the least-sophisticated investors are the most likely to prefer that their foreign exchange trades be executed at the precise time at which the benchmark is fixed (Foreign Exchange Benchmark Group 2014).[2]

In the special case of interbank offered rates, there is an important additional motive for introducing a benchmark. Suppose a bank wishes to hedge the risk of a change in its borrowing cost. However, because the bank is known to have private information about its idiosyncratic credit quality, it might find that no counterparty is eager to hedge this risk. This problem of adverse selection can make it hard for the bank to negotiate a contract that is based on its own future credit spread. This market breakdown might be overcome to some extent with hedging contracts that are instead linked to market-wide, rather than bank-specific, credit spreads. In this case, a benchmark based on the interbank offered rate allows banks to hedge at least the common component of their borrowing costs.

Agglomeration of Trade around Benchmarks

Once a benchmark has been established, it can become a powerful "basin of attraction" for related trades, based on two types of agglomeration effects. To see why, suppose that a spectrum of possible non-benchmark trades could be substituted with a benchmark trade. These alternative types of trades are differentiated by their risk attributes and other characteristics, such as time of execution (relative to the time at which the benchmark is fixed).

One force driving agglomeration is the incentive for market participants to reap the information-related benefits of a benchmark that we described in the previous section, including lower search costs, higher market participation, better matching efficiency, and lower moral hazard in delegated execution. In order to obtain these benefits, market participants or their agents will often choose to substitute their "best-fit-for-purpose" trade with a benchmark trade. For example, a foreign-exchange trade that, absent benchmark effects, would optimally be executed at 5 pm London time could be shifted to match the extremely popular WM/Reuters benchmark, produced by the WM Company, which has a 4 pm London fixing time. Similarly, an investor who is interested in taking a hedging or speculative position in risk-free interest rates might shift toward a LIBOR-based financial instrument, even though the bank-credit-spread component of LIBOR is somewhat undesirable.

The second force for agglomeration is the incentive to lower trading costs that are associated with illiquidity. A high volume of trade in a financial instrument is typically associated with a smaller bid–ask spread, shorter execution delays, lower search costs, and a lower price impact for large trades. Once trading in a benchmark-related instrument is active, there is an incentive to substitute from

[2] Indeed, less-sophisticated investors in foreign exchange markets commonly request "fix trades," by which they contract with a dealer to buy or sell at the benchmark price itself, without a fee or bid–ask spread. The dealer absorbs the risk of laying off the position acquired from its clients at a different price, and thus a potential loss. The dealer may be compensated in part, however, by the common practice in this market of "front running" by dealers, who may trade on their own behalf a few seconds before the fixing, thus causing a price impact to the fixing that can benefit the dealer at the expense of its clients. Whether malicious or not, the report commissioned by the Financial Stability Board recommends that this practice be curtailed.

less-actively traded instruments toward instruments that reference the benchmark (McCauley 2001). This liquidity incentive can easily dominate any mildly undesirable investment characteristics of a non-benchmark instrument.

Once a benchmark is established, its basin of attraction can thus become larger and larger, given the positive feedback effects of informational transparency and liquidity. In the next section, we provide some statistics that illustrate the extent to which LIBOR has become the overwhelmingly popular interest-rate benchmark.

Once liquidity in LIBOR-linked contracts became firmly established in the 1980s, dealers and derivatives exchanges had the incentive to introduce a wide range of LIBOR-based hedging instruments, including exchange-traded eurodollar futures and options available from Chicago Mercantile Exchange Group, and over-the-counter derivatives including caps, floors, and swaptions (that is, an option to engage in a swap contract). The availability of risk transfer in these related instruments further increased the magnetic qualities of LIBOR-based trading.

Manipulation and Manipulation-Resistant Fixing Methods

During the financial crisis of 2007–2009, no bank wished to appear to be less creditworthy than others, as concerns over their creditworthiness might have raised their costs of funding, or in the extreme case, caused a run. When banks were polled to produce LIBOR, the rates reported by each bank were listed individually. As a result, some banks started “low-balling”—that is, understating their true borrowing costs when submitting to the LIBOR poll. The unrealistically tight bunching among banks of their reported borrowing rates is part of what led to the news reports of likely manipulation. Subsequent research revealed a substantial downward and persistent bias in LIBOR relative to actual bank borrowing rates (Abrantes-Metz, Kraten, Metz, and Seow 2012; Snider and Youle 2012; Kuo, Skeie, and Vickery 2012; for an overview, see Hou and Skeie 2013).

The second basic motive for manipulating benchmarks is a desire to profit on positions in derivative financial instruments that are contractually linked to the benchmark. In the case of interbank offered rates, some derivatives traders asked bank officials that were charged with providing rate submissions to the LIBOR poll to bias their reports. Figure 1 offers some examples of emails between traders that later emerged in an investigation of Barclays Bank. Sometimes these requests would be relayed by another trader, often located at another bank. In some instances, more significant distortions were achieved through collusion that coordinated the misreporting among several banks.

Clearly, if traders are able to benefit their swap positions by causing a benchmark to move one way or the other, the least ethical of them may attempt to do so. The extent to which a fixing can be distorted will always be a source of incentive to manipulate. However, an additional incentive is the ease with which very large positions in LIBOR-linked derivatives can be established, given the extremely high volumes and liquidity in this market.

In order to mitigate manipulation, tighter governance and regulatory monitoring of the fixing process may be somewhat effective, especially for those benchmarks that are set by judgment-based reporting (United Kingdom Financial

Figure 1
Some Emails and Text Messages from Swaps Traders at Barclays

1) "WE HAVE TO GET KICKED OUT OF THE FIXINGS TOMORROW!! We need a 4.17 fix in 1m (low fix) We need a 4.41 fix in 3m (high fix)" (November 22, 2005, Senior Trader in New York to Trader in London).

2) "You need to take a close look at the reset ladder. We need 3M to stay low for the next 3 sets and then I think that we will be completely out of our 3M position. Then it's on. [Submitter] has to go crazy with raising 3M Libor." (February 1, 2006, Trader in New York to Trader in London).

3) "Your annoying colleague again. … Would love to get a high 1m Also if poss a low 3m … if poss. … thanks" (February 3, 2006, Trader in London to Submitter).

4) "This is the [book's] risk. We need low 1M and 3M libor. PIs ask [submitter] to get 1M set to 82. That would help a lot" (March 27, 2006, Trader in New York to Trader in London).

5) "We have another big fixing tom[orrow] and with the market move I was hoping we could set the 1M and 3M Libors as high as possible" (May 31, 2006, Trader in New York to Submitter).

Source: From the investigation of Barclays by the US Commodity Futures Trading Commission, http://www.cftc.gov/ucm/groups/public/@lrenforcementactions/documents/legalpleading/enf barclaysorder062712.pdf.
Notes: The references to 3m or 1m refer to three-month or one-month LIBOR estimates. The term "fix" refers to the actual LIBOR announcement on a given day.

Conduct Authority 2012). But the first line of defense is having a benchmark definition and a fixing methodology that are more difficult to manipulate.

All else equal, it is better to have the benchmark fixing based on a large volume of transactions so that it is difficult for individual manipulated trades or reports to have much influence on the fixing and so that it is easier to detect when trades and reports are distortionary. This can be achieved in part by widening the time window over which rates or prices are averaged to determine the benchmark and by broadening the set of instruments or types of trades that are used. Specific recommendations for broadening the data collected to fix benchmarks have been made for the interbank offered rates (Market Participants Group on Reforming Interest Rate Benchmarks 2014; Official Sector Steering Group 2014; Duffie, Skeie, and Vickery 2013) and for the foreign exchange benchmarks (Foreign Exchange Benchmark Group 2014).

A key tradeoff is that broadening the data collected to fix a benchmark can increase the heterogeneity of the proxies used for the item being measured, whether through timing or quality differences. This heterogeneity can be mitigated with statistical methods, but in the end the benchmark may be more robust to manipulation but not very specific to the trading interests of market participants. One way to do better is to weight the data strategically so as to produce a fixing that efficiently trades off the incentive to manipulate against measurement error. For example, smaller trades (those whose prices are most easily distorted) are optimally downweighted (Duffie and Dworczak 2014).

Table 1
US Dollar LIBOR Market Footprint by Asset Class and Tenor

Asset class	*Volume (billions of dollars)*	*% LIBOR-related*	*Most common tenors (in months)*
Loans			
Syndicated loans	~3,400	97%	1m and 3m
Corporate business loans	1,650	30-50%	1m and 3m
Noncorporate business loans	1,252	30–50%	1m and 3m
Commercial real estate/Commercial mortgages	3,583	30–50%	3m
Retail mortgages	9,608	15%	6m
Credit cards	846	Low	
Auto loans	810	Low	
Consumer loans	139	Low	
Student loans	1,131	7%	1m and 3m
Bonds			
Floating/Variable Rate Notes	1,470	84%	1m and 3m
Securitizations			
Residential Mortgage-Backed Securities (RMBS)	~7,500	24%	1m (83%)
Commercial Mortgage-Backed Securities (CMBS)	~636	4%	1m (75%)
Asset-Backed Securities (ABS)	~1,400	37%	1m (76%)
Collateralized Loan Obligations (CLO)	~300	71%	3m (82%)
Over-the-counter derivatives			
Interest-rate swaps	106,681	65%	3m (90%)
Forward Rate Agreements (FRAs)	29,044	65%	3m (90%)
Interest-rate options	12,950	65%	3m (90%)
Cross-currency swaps	22,471	65%	3m (90%)
Exchange-traded derivatives			
Interest-rate options	20,600	98%	3m
Interest-rate futures	12,297	82%	3m

Source: This table is adapted from *Market Participants Group on Reforming Interest Rate Benchmarks, Final Report*, March 2014.

Reforming LIBOR

How is LIBOR Used?

With this general framework in mind, we now turn to the specific problem of reforming LIBOR. Most of the issues that we will discuss pertain to all of the LIBOR currencies—US dollar, British pound, euro, Swiss franc, and Japanese yen—as well as to the other IBORs, namely EURIBOR and TIBOR. For concreteness, we focus on the case of the US dollar LIBOR.

Table 1 presents some facts concerning the major applications of US dollar LIBOR, drawn from Market Participants Group on Reference Rate Reform (2014). The table covers four broad categories of financial instruments: loans, bonds, securitizations, and derivatives—both over-the-counter and exchange-traded. Several points stand out. First, across a range of applications, a majority of contracts tend to be linked to either the 1-month or 3-month LIBOR rate. Second, LIBOR is the dominant interest-rate benchmark for trillions of dollars of conventional loans, many of which are retained on the balance sheets of banks and other intermediaries. For example, 97 percent of the $3.4 trillion syndicated loan market is tied to LIBOR. Among other business loans and commercial real estate loans, which collectively add up to nearly $6.5 trillion, somewhere between 30 and 50 percent are linked to LIBOR. Of the $9.6 trillion of nonsecuritized residential mortgages, about 15 percent have adjustable interest rates that are LIBOR-based.

For lending applications that appear on bank balance sheets, it is easy to understand the appeal of an interest-rate benchmark like LIBOR that embeds an element of bank credit risk. If a commercial bank makes a long-term floating-rate business loan or commercial real estate loan, and funds the loan by borrowing short-term in the wholesale unsecured market, the bank's funding costs are exposed to movements in both the general level of interest rates as well as the bank's credit spreads. Thus, if the floating-rate loan is tied to an index based on a riskless rate, like the Treasury bill rate, then the bank has hedged only the component of its funding costs that is related to riskless rate. If credit spreads for the banking industry widen *relative to the riskless rate*, the bank's net interest margin will suffer. Indeed, during the 2007–2009 crisis, LIBOR rates went up several percentage points, whereas Treasury rates declined! If the floating-rate loan is instead linked to LIBOR, then the bank will at least be hedged with respect to the market-wide component of bank credit spreads, albeit not to idiosyncratic movements in its own credit spread. As discussed earlier, this motive for hedging risks that appear on bank balance sheets helps to explain why early efforts at creating interest-rate benchmarks in the 1970s gravitated toward a rate like LIBOR that was intended to capture bank credit risk.

But sizable and important as these bank-related lending applications are, the most striking fact in Table 1 is how they are now utterly dwarfed by trade in interest-rate derivatives tied to LIBOR. For example, the dollar-based over-the counter interest-rate swap market alone is estimated to be on the order of $107 trillion in gross notional value, of which 65 percent is linked to LIBOR. (In this market, an investor who prefers to pay a fixed interest rate rather than a variable rate such as 3-month LIBOR can enter a swap that exchanges the difference between these rates, for a given number of years, with another market participant that has the opposite preference.) Roughly another $100 trillion in interest-rate derivatives—including futures, cross-currency swaps, and both over-the-counter and exchange-traded options—are heavily LIBOR-dependent.

In contrast to the use of LIBOR for hedging a bank's loan funding costs, it is improbable that many users of interest-rate derivatives have an intrinsic economic reason to be exposed to the changes in bank credit spreads that are included in LIBOR. Rather, the majority are likely using these derivatives either to hedge an exposure to the general level of interest rates, to make a speculative bet on market-wide interest rates, or to intermediate such trades. For these "rates traders," the fact that LIBOR incorporates a bank credit risk component is, if anything, a bit of a nuisance. This inconvenience is apparently more than offset by the liquidity advantages of trading in the tremendously deep LIBOR-based derivatives market, as discussed earlier.

Costs of Excessive Agglomeration around the LIBOR Benchmark

In the narrative that we have in mind, bank-hedging motives were the seed that originally made LIBOR an attractive benchmark. From this seed, and given the strong agglomeration effects associated with liquidity and transparency externalities, the market for interest-rate-linked products has grown exponentially, while the benchmark has remained "stuck" on LIBOR. This is so despite the fact that much of the subsequent demand for referencing an interest-rate benchmark has come from users—most notably derivatives traders—who care a great deal about liquidity and transparency but who may have no particular desire for exposure to the bank-credit-risk component of LIBOR.

If this story is correct, it suggests that two distinct costs are associated with the pileup of so much trading on LIBOR-linked contracts. First, LIBOR may offer a less-than-ideal fit for the purposes of the majority of derivatives users. That is, even if most derivatives users would prefer to have their contracts tied to another benchmark without a bank credit risk component (for example, Treasury bills), once LIBOR has become the dominant benchmark, it is very difficult for the market to switch to this new equilibrium on its own. The result of these liquidity externalities can be that markets suffer a coordination failure and become stuck at an inferior equilibrium.

Second, the incentives for manipulation are heightened when a large derivatives market is indexed to a benchmark rate that is set in a primary market where trading activity is orders of magnitude smaller. What is striking about many of the documented cases of LIBOR manipulation is that they involved only very small rate distortions, with the guilty parties often misstating their borrowing costs by just a few basis points. Even such tiny distortions in LIBOR fixings can be potentially very profitable for a manipulator who has accumulated a large enough position in derivatives whose payments are contractually based on the LIBOR fixing. Thus the relative scales of the two markets—the derivatives market versus the primary market which ultimately determines the reference rate—play a key role in manipulation incentives.

Moreover, this manipulation problem is not resolved merely by improving the design of the LIBOR fixing methodology, despite the importance of making these improvements. In the past, manipulators arranged for dishonest judgment-based reports of bank borrowing rates. But even with a fully transactions-based benchmark, a manipulator might attempt to distort actual transactions in the underlying bank funding markets. A thin underlying borrowing market cannot be expected to provide a robust foundation for a multi-hundred-trillion dollar derivatives market, even with substantial improvements to the LIBOR fixing methodology.

The Basic Idea of a Two-Benchmark Approach

If we were starting from scratch, what might a more efficient and resilient set of arrangements for interest-rate benchmarking look like? The above discussion suggests that there could be considerable appeal in a "two-rate approach," that is, two distinct types of interest-rate benchmarks. One of these, an improved version of LIBOR itself, would continue to be based on banks' wholesale unsecured funding costs and would be appropriate for applications that rest on that credit risk component, such as hedging the revenues of balance-sheet lenders. This banking-oriented benchmark would be reformed so as to be transactions-based and subject to a tougher monitoring regime, and hence less subject to manipulation.

The second benchmark would be based on a riskless or near-riskless rate that is established in a broad and deep market. The goal here would be to give pure interest rate traders—potentially a large fraction of the derivatives market—something that fits their risk-transfer needs well, while at the same time reducing the manipulation incentives that arise when so much rates-trading is tied to a rate like LIBOR that is based on the much thinner underlying market for unsecured bank borrowing.

For the two-rate approach to be more fully articulated, three questions need to be addressed. First, how does one most effectively design an improved version of LIBOR, which we will call LIBOR+, so that it is based to the maximum extent possible on actual market transactions, rather than on banks' discretionary reports of their funding costs? Second, what is the appropriate riskless or near-riskless rate to use for pure rates-trading applications? Third, and perhaps most challenging, given that we are not actually starting from scratch, and given the large obstacles posed both by legacy contracts and liquidity-driven coordination problems, how can policymakers help to break the stranglehold of existing LIBOR and pave the way for transition to a two-rate regime? In what follows, we consider each of these questions in turn.

The Design of LIBOR+

The various policymaking groups that have studied the manipulation problems associated with LIBOR have all concluded that it would be desirable to move away from the current practice of fixing LIBOR rates based on judgmental submissions from a panel of banks and shift to a fixing methodology that is more anchored in observable, verifiable market transactions. In addition to whatever benefits such a switch might bring in terms of reduced manipulability, if the fixing methodology is entirely algorithmic, it would also eliminate a potential threat to financial stability—namely that, because of legal risks, member banks might decide to defect from the LIBOR panels, making it impossible to calculate a reliable reference rate under the poll-based methodology. In the case of EURIBOR, the euro-based interbank offered rate, there has already been a notable exodus from the panel of reporting banks, which had dropped from a high of 44 to only 26 banks by June 2014 (Brundsen 2014). With an algorithmic approach to fixing, there is no need for banks to decide whether they will contribute to a LIBOR panel.

Although a transactions-based approach has clear appeal, it is more difficult to implement than one might first think. For example, 3-month LIBOR is meant to reflect the typical rate at which large banks borrow on an unsecured basis for a 3-month term *from other banks*. But the volume of borrowing in the interbank market is small and has been secularly trending downward. Some of the secular decline in interbank borrowing is likely due to the extraordinary monetary policies of the last several years, which have left banks glutted with reserves and therefore less dependent on interbank borrowing to manage their liquidity positions. There is

Table 2
Transactions Data on Unsecured Bank Borrowing

		Number of Trades					Numbers of Issuers					Volume ($mn)				
		O/N	1W	1M	3M	6M	O/N	1W	1M	3M	6M	O/N	1W	1M	3M	6M
Daily Avg	2014	468	74	21	19	18	15	9	7	8	7	20,223	3,204	888	706	718
	2013	511	95	18	25	13	16	9	6	8	6	22,312	4,157	702	1,006	474
	2012	344	62	24	31	13	17	10	8	9	5	14,889	2,637	888	1,211	452
	2011	435	79	38	34	18	21	15	14	11	5	18,945	3,356	1,407	1,331	706
Daily Max	2014	538	127	42	45	40	17	13	10	12	11	23,853	5,460	1,869	1,903	1,861
	2013	878	280	78	126	76	20	18	13	17	15	39,722	13,043	3,479	5,904	2,892
	2012	521	225	80	112	55	24	20	19	19	13	22,985	10,007	3,613	4,539	2,140
	2011	666	263	113	107	112	27	25	32	24	15	30,015	11,686	4,982	4,642	4,985
Daily Min	2014	406	31	3	8	2	14	5	3	4	2	16,998	1,279	77	222	50
	2013	187	7	1	1	1	13	3	1	1	1	6,910	204	5	1	1
	2012	33	4	0	2	0	7	2	0	1	0	1,399	124	0	64	0
	2011	235	10	3	3	0	17	4	1	1	0	9,608	242	75	24	0

Source: Market Participants Group, *Final Report*, March 2014.
Notes: This table displays daily average, maxima, and minima for number of trades, number of issuers, and dollar volume of unsecured bank borrowing transactions in the commercial paper (CP) and certificate-of-deposit (CD) markets based on a sample from a unit of J.P. Morgan over the period 2011 through January 2014. Maturity buckets are defined as follows: O/N = 1 day to 4 days, 1W = 6 days to 8 days, 1M = 28 days to 32 days, 3M = 85 days to 95 days, 6M = 175 days to 185 days. "$mn" means "millions of dollars."

also a significant reduction in interbank unsecured borrowing during periods of market stress. This tendency is especially pronounced at longer maturities (Gorton, Metrick, and Xie 2014). The paucity of interbank lending is especially severe in Japanese yen and Swiss francs (Market Participants Group on Reforming Interest Rate Benchmarks 2014).

Simply put, most banks don't borrow at longer maturities from other banks on most days. This is an obvious challenge to any attempt to measure term interbank borrowing rates on a daily basis, be it judgment-based or transactions-based. If LIBOR is to serve as an effective benchmark, its fixing should be broadened so as to be based on unsecured bank borrowings *from all wholesale sources*—not just other banks, but nonbank investors in bank commercial paper and large-denomination certificates of deposit (CDs). This is a key recommendation for LIBOR+ in the Market Participants Group (2014) report. Indeed, this report conducted a pilot study of LIBOR+ using proprietary data from a unit of J.P. Morgan that covers approximately 40 to 45 percent of the overall market for unsecured bank borrowing. Table 2 gives some details on the density of transactions at various tenors (that is, lengths of borrowing period) in this data set. For example, over the period from 2011 to early 2014, there were roughly 25 to 30 transactions at the 3-month tenor on a typical day, for an average total daily dollar volume of about $1 billion. However, on the lowest-volume days, there were only a handful of transactions, numbering on the order of three to eight.

With these data in hand, the Market Participants Group (2014) built a prototype LIBOR+ fixing algorithm. Their basic methodology is as follows. On any given day t, for any given bank i, and for any tenor of interest, if bank i has an available transaction, the rate on that transaction is entered with a weight of unity into the index. If bank i does not have an available transaction, the algorithm goes back to the nearest prior date $t - k$ when there is a transaction, and enters the rate on that transaction into the index with a reduced weight—one that gets smaller as the distance k from the present gets larger. Thus the algorithm includes noncontemporaneous data to compensate for the low density of transactions on any given day, but downweights the older data in light of its staleness (Duffie, Skeie, and Vickrey 2013).

The results of this exercise are plotted in Figure 2, which compares the constructed LIBOR+ to actual LIBOR for each of the 1-, 3- and 6-month tenors. As can be seen, while LIBOR+ is always more volatile on a day-to-day basis than LIBOR—which is not surprising given the opinion-based nature of LIBOR—the levels of the two time series track each other reasonably closely at both the 1-month and 3-month tenors. At the 6-month tenor, the fit is considerably less good. Some of this deterioration in fit is due to the paucity of transactions at 6-month terms. But some of it is due to a particular form of sample selection—the fact that during a period of market stress, only the highest credit-quality banks find it economically sensible to issue at a 6-month maturity. This selection effect tends to make the transactions-based LIBOR+ lower than the judgmentally reported LIBOR during stressful periods in the banking sector. Nevertheless, given that the vast majority of contracts in dollar LIBOR reference the 1- and 3-month tenors, the LIBOR+ methodology holds considerable practical promise, especially if the interbank loan data supporting it can eventually be augmented to capture the entire universe of certificate-of-deposit and commercial paper transactions.

However, even if a transactions-based LIBOR+ methodology can be made to work well from an economic perspective, there remains the crucial question of whether it also "works" legally. In other words, for the large stock of existing legacy contracts that reference LIBOR, is it possible to seamlessly substitute a fixing along the lines of LIBOR+ without causing private litigants to challenge this substitution? We will return to this question later.

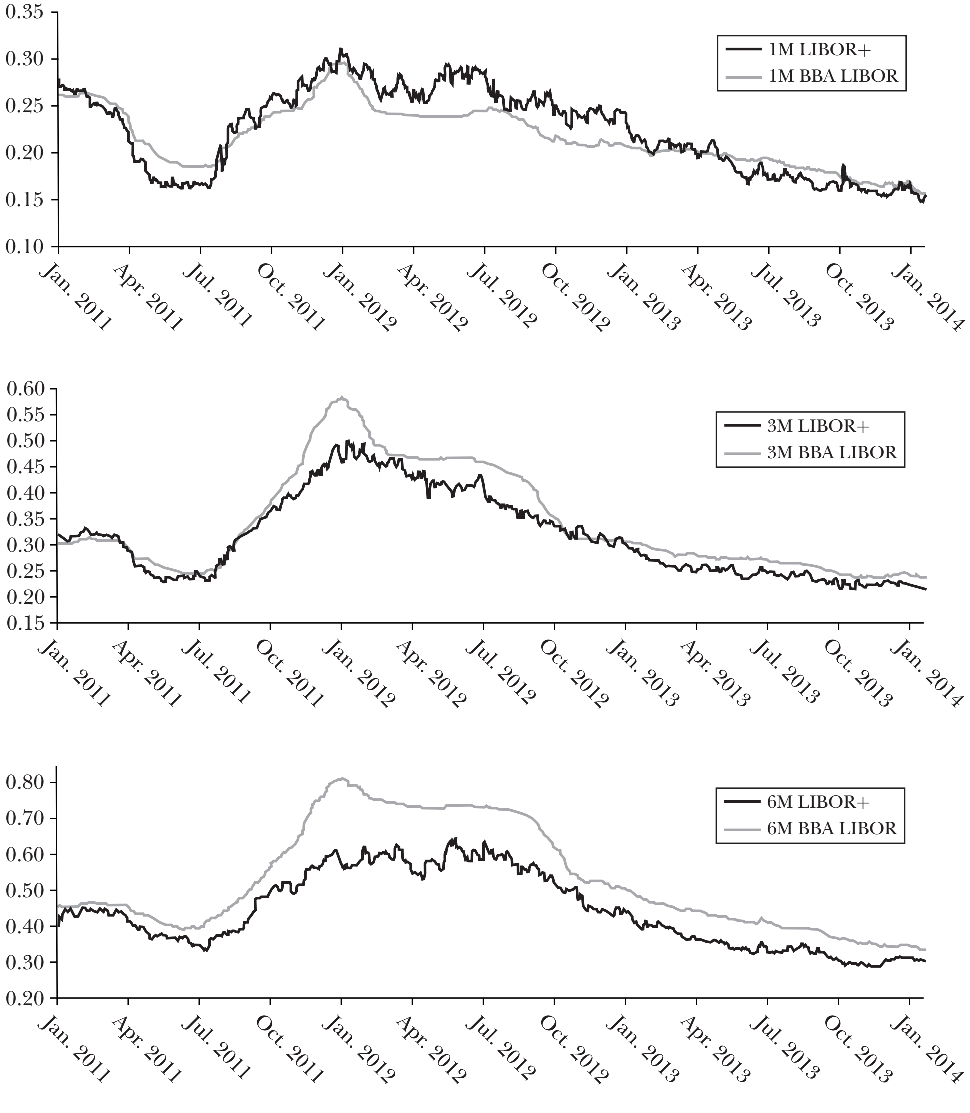
What is a Suitable Riskless Interest Rate Benchmark?

Despite the potential promise of LIBOR+ for certain bank-based transactions, we believe that it would be a mistake for such a benchmark to shoulder the burden of being the primary reference rate for the entire interest-rate derivatives market. To understand why, compare the magnitudes in Tables 1 and 2. At the commonly used 3-month tenor, transactions in the underlying market for unsecured bank funding are roughly on the order of \$1 billion dollars on a typical day, while the volume of gross notional outstanding in the swap market that references LIBOR at this tenor is on the order of \$100 trillion, or *100,000 times larger.*[3] As we have been emphasizing, this divergence leaves a strong incentive for a trader with a large derivatives position to manipulate even a transactions-based LIBOR+, for example

[3] This compares a flow with a stock, but the difference remains striking. The daily volume of US dollar LIBOR-based derivatives has an order of magnitude of around \$1.15 trillion (Bank for International Settlements 2013, table 3). This is roughly a factor of 1,000 times the volume of trade determining 3-month US dollar LIBOR. Moreover, payments on the much larger stock of outstanding derivatives are exposed to daily LIBOR fixings.

Figure 2

Comparison of Transactions-Based LIBOR+ to Actual British Banker's Association (BBA) LIBOR



Source: Market Participants Group (2014).

by borrowing or lending (or by arranging for someone else to borrow or lend) more or less aggressively in the markets for interbank loans, commercial paper, or certificates of deposit.

In our view, a key to reforming interest rate benchmarks is therefore to encourage the transition of a large fraction of derivatives trading to a more robust benchmark based on interest rates that are risk-free, or nearly so. There are several possibilities for such a benchmark: an interest rate administered by the central bank, the rate on short-term Treasury bills, general collateral repo rates, and overnight index swap (OIS) rates. We consider each of these in turn.

The Federal Reserve sets certain interest rates directly. For example, it sets the rate that it pays to banks on their excess reserves. It also sets the "overnight reverse repurchase rate," which is the rate paid by the Fed to a wider range of market participants on overnight reverse repurchase agreements, whereby the Fed effectively borrows on a collateralized basis against its holdings of government securities. Indeed, the Fed has announced that it plans to use its control of these two rates as tools to implement changes in monetary policy going forward.

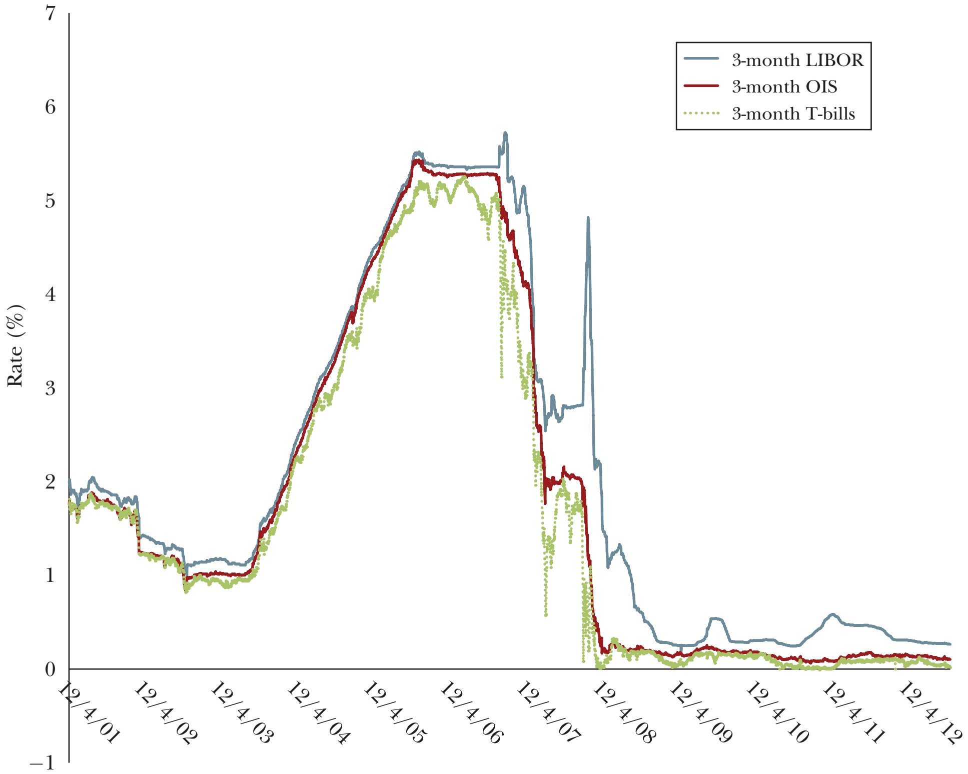
Because these two rates are directly administered by the Fed, as opposed to being set in the market, they are immune from manipulation. However, the appeal to market participants of using these administered rates as benchmarks will depend on the details of how the Fed uses them to implement monetary policy. For example, the Fed may decide to leave a relatively large spread between the rate on excess reserves and the reverse repurchase rate, with market-determined short-term rates bouncing between these two administered rates. In that case, neither of the two rates would be a tight proxy for the market risk that investors are most anxious to transfer. A secondary concern is whether an overnight interest rate like that on the reverse repurchase facility would be attractive for the settlement of floating-rate contracts that have traditionally been based on longer maturities such as three months.

The rate on short-term Treasury bills is another natural candidate for a riskless reference rate. While this market is not manipulation-proof, it is certainly much deeper and more active than the market for unsecured bank borrowing. Moreover, in January 2014, the US Treasury began to issue floating interest rate notes linked to auction-determined rates on 13-week Treasury bills. The Treasury's presence in the floating-rate note market may help to boost liquidity in contracts that use Treasury bill interest rates as a benchmark.

The Market Participants Group (2014) report received input from a wide range of market participants regarding their desire to use Treasury bill rates as a reference rate for derivatives contracts. The responses indicated a general lack of enthusiasm for this option. One reason for this skepticism is that during periods of market stress, "flight-to-quality" or "safe-haven" demands tend to lower the rates on Treasury bills relative to those on other relatively safe instruments. This phenomenon is illustrated in Figure 3, which plots the rate on 3-month Treasury bills along with the 3-month overnight index swap (OIS) rate, another often-used proxy for a near-riskless rate (which will be discussed further later in this section), as well as 3-month LIBOR. Several downward spikes of the Treasury bills rate relative to the OIS rate during the financial crisis are readily apparent. To the extent that investors are seeking to hedge or speculate on the general level of rates without taking a view on movements in these Treasury-bill-specific safe-haven premiums, these spikes can make the interest rate on Treasury bills less appealing as a reference rate.

Nevertheless, we think it is easy to exaggerate this concern. Over the sample period December 2001 to July 2013 shown in Figure 3, the correlation between 3-month Treasury bill rates and 3-month OIS rates is 0.995 in levels and 0.560 in

Figure 3
3-Month LIBOR, Overnight Index Swap (OIS), and Treasury Bills



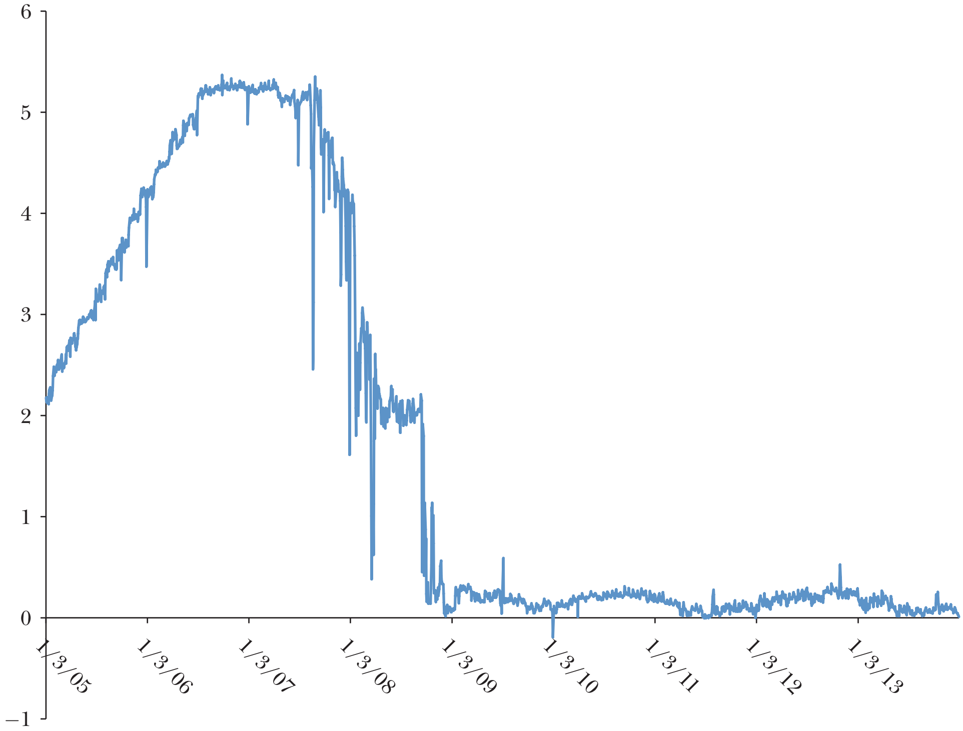
Source: Data is from Bloomberg.

weekly changes. The basis risk here is notable mainly in tail events. Given the other obvious merits of using Treasury bills as a reference rate, our view is that this option should be given careful consideration.

Another near-riskless rate is the so-called "Treasury general collateral repo rate." A "general collateral" repurchase agreement is signed without specifying a particular security that will be sold and repurchased, but instead just specifying that the lender of funds will accept anything from the general class of Treasury and other related securities as collateral. Thus, the general collateral repo rate is effectively the average rate at which dealers obtain overnight financing secured by Treasury securities. This market is highly liquid; recently, about $590 billion of Treasuries are financed this way on a typical day.[4] Hence, like Treasury bill rates, one would expect general collateral repo rates to be relatively robust to manipulation.

[4] The Federal Reserve Bank of New York provides the amounts of securities financed in the tri-party repo market on the seventh business day of each month. For July 2014, see http://www.newyorkfed.org/banking/pdf/jul14_tpr_stats.pdf.

Figure 4
Overnight Treasury General Collateral Repo Rate



Source: The data is from Boomberg.
Notes: The data in the figure is for the Treasury General Collateral Finance (GCF) rate, which is published by the Depository Trust and Clearing Corporation. The GCF rate is based on a subset of transactions (approximately 20 percent) in the broader tri-party general collateral repo market.

Although there is currently no official general collateral repo rate, Figure 4 plots a close proxy, the "Treasury General Collateral Finance" rate, which is published by a company called the Depository Trust and Clearing Corporation. This rate is based on a subset of about 20 percent of all transactions in the broader general collateral repo market. As shown, general collateral repo rates, like Treasury bill rates, tend to spike downward during periods of market stress, reflecting a safe-haven property. Some of the volatility of the general collateral repo rate is also due to the one-day maturity of this rate. That is, unlike the 3-month Treasury bill rate, there is no "averaging out" of the impact of short-lived supply and demand shocks. A further disadvantage of the general collateral repo rate is that the underlying market is not very active at maturities beyond one week, whereas LIBOR is most actively referenced at the 1-month and 3-month maturities.

Motivated by these limitations with the general collateral repo rate, a more novel benchmark design discussed in the Market Participants Group (2014) report is the compounded interest rate implied by the overnight general-collateral rates over (say) the three months leading up to settlement date.[5] This 3-month-lagged compounded daily rate is clearly an implementable benchmark. This rate is extremely robust to manipulation because, as we have discussed, the underlying general collateral repo rate is itself robust, and the averaging inherent in this formula makes manipulation all the more difficult. On the other side of the ledger, a potential drawback of this compounded-daily-rate benchmark is its backward-looking nature. Still, while some market participants might prefer to know their next floating-rate interest payment well in advance of the due date rather than waiting until very shortly before the payment is due, this wait-and-see payment method is more familiar to most wholesale market participants such as swaps traders. Even retail financial consumers are familiar with the idea of having their latest floating-rate mortgage payments reported to them after the fact in their bank statements, in the same manner as their utility payments.

Our final candidate for a low-risk interest rate benchmark, as we mentioned earlier, is the overnight index swap (OIS) rate. The 3-month OIS rate is the interest rate on a so-called overnight index swap, which pays a predetermined fixed interest rate in exchange for receiving the compounded daily federal funds rate over the 3-month term of the contract.[6] Thus, the 3-month OIS rate can be thought of as the market's forward-looking expectation for the average federal funds rate that will prevail over the upcoming three months. (Because of compounding and also because of risk aversion to uncertain changes in future daily federal funds rates, this "expectation" is slightly biased.) To the extent that federal funds interest rate transactions—which are overnight unsecured borrowings by banks—are themselves close to riskless, the 3-month OIS rate is a reasonable proxy for a 3-month riskless rate. An advantage of OIS is that it does not incorporate the same kind of safe-haven premium as Treasury bills.

The potential appeal of the overnight index swap rate as a standardized low-risk rate is evident in Figure 3. During periods of market stress, there are no upward spikes associated with jumps in term credit risk premiums, and no downward spikes associated with flight to a Treasury-like safe haven. Some researchers and many market practitioners therefore rely on OIS rates as a relatively clean and transparent proxy for the "true" riskless rate.

However, it is less clear that the overnight index swap rate is ready for the more demanding task of serving as a benchmark for payments on many trillions of dollars of interest-rate derivatives. Importantly, the OIS market itself is a derivatives market that is not yet heavily traded. For example, Fleming, Jackson, Li, Sarkar, and Sobel (2012) report that between June and August 2010 there were an average of only 31 transactions a day in US dollar OIS, representing a notional volume of about $30 billion. It is not clear that one should attempt to shift volume from a LIBOR

[5] For example, the contractually agreed floating-rate payment due at the end of a 90-day contract period would be $P = (1 + r_1)(1 + r_2)\ldots(1 + r_{T-1})$, where $T = 90$ and where r_k is the overnight general collateral repo rate.

[6] More generally, by entering an overnight index swap position as a fixed-rate payer, one agrees to pay at maturity in T days the notional principal amount plus interest on this amount at the contractually agreed

benchmark on the premise that the underlying bank-borrowing market is so thin, and then substitute with another rate such as OIS that is also set in a relatively thinly traded market.

None of the alternative reference rates that we have discussed is perfect for all applications, but they are feasible and relatively effective substitutes for many applications currently served by LIBOR. None of these alternative rates include a significant component for bank credit risk, which is an advantage over LIBOR for most "rates trading" applications. All of these alternatives, with the exception of the overnight index swap rate (whose market is currently relatively thin), are far less subject to manipulation than LIBOR. If the OIS market were to grow sufficiently, perhaps boosted by support from the official sector, it too might someday become an effective substitute for a significant amount of LIBOR-based derivatives trading, though in our judgment it is not currently up to the task.

Can We Get There from Here? Transition Challenges

To summarize the discussion to this point: We have argued that if we could start the world from scratch, we would aim for a two-rate model, with a transactions-based LIBOR+ serving as the reference rate for most on-balance-sheet bank lending contracts, and with some low-credit-risk reference rate—such as the Treasury bill rate, the 3-month lagged compounding of daily general collateral repo rates, or perhaps eventually the overnight index swap rate—serving as the reference rate for the majority of interest-rate derivatives. However, given the large stock of legacy contracts already tied to LIBOR, as well as the strong liquidity-driven network effects that we have discussed, getting from here to there presents formidable challenges. In what follows, we lay out a transition strategy that we think has the best shot of addressing these challenges. We acknowledge, however, that even this best-case strategy faces a number of daunting uncertainties.

A "Seamless" Transition from LIBOR to LIBOR+ for Legacy Contracts

The first element in our idealized transition strategy is a "seamless" transition from LIBOR to LIBOR+ for legacy contracts. At some future date, the LIBOR administrator would stop publishing LIBOR based on its current fixing methodology, and would begin publishing LIBOR+ in its place. The current administrator for LIBOR is ICE Benchmark Administration, which took over from the British Bankers' Association (BBA) in early 2014. Contracts would not need to be rewritten to change the stated "LIBOR" reference rate; merely the fixing would change.

The key risk associated with this approach is that it may provoke legal challenges, in which one party to a contract claims that his obligations should be

OIS rate R, in exchange for a floating-rate payment from the counterparty. The floating-rate payment per dollar of notional is the compounded overnight amount, that is, $(1 + r_1)(1 + r_2)\dots(1 + r_{T-1})$, where r_k is the stipulated benchmark overnight interest rate on day k.

discharged based on the doctrine of "contract frustration." The Market Participants Group (2014) report contains a detailed analysis of this issue. Although it is hard to be definitive, the report suggests that contract-frustration risks are likely to be mitigated if: 1) the conceptual basis for LIBOR+ (as a rate representative of unsecured bank borrowing costs) is close to that for existing LIBOR; and 2) the two rates have similar levels as of the transition date, as well as similar statistical properties, thereby minimizing any value reallocation associated with the switch. The report concludes, based on legal consultations as well as the sort of pilot-testing of LIBOR+ shown in Figure 2, that a "seamless transition can be achieved for US dollar LIBOR at the popular 1-month and 3-month tenors without raising undue risk of legal contractual frustration risk" (p. 25). However, the report does not reach a final conclusion about whether to attempt a seamless transition from LIBOR to LIBOR+ at the somewhat less-popular 6-month maturity.

Pushing Newly Written Derivatives to a Riskless Reference Rate

While a seamless transition appears to be a promising approach for moving contracts from LIBOR to LIBOR+, it is unlikely to be viable for moving contracts from LIBOR to an alternative low-credit-risk rate of the sort we have discussed, for example the Treasury bill rate. The differences between LIBOR and these other alternatives are too substantial, both in concept and in statistical behavior, for such a switch to avoid legal challenges based on contract frustration. Instead, if the goal is to move a major fraction of derivatives trades to a riskless rate, this must be accomplished differently. We propose the following steps.

First, the majority of already-existing derivatives contracts would not be altered, but rather could simply be allowed to roll off over time. An analysis of the maturity distribution of these contracts suggests that a substantial "roll-off" can occur over a five-year horizon. Specifically, for the different categories of over-the-counter and exchange-traded derivatives displayed in Table 1, about 65 percent of over-the-counter interest-rate swaps will roll off over five years, as would approximately 100 percent of floating-rate agreements, 74 percent of over-the-counter interest-rate options, 76 percent of cross-currency swaps, 100 percent of exchange-traded interest-rate options, and 99 percent of exchange-traded interest-rate futures (Market Participants Group 2014, p. 309).

Second, regulators would use a variety of tools to encourage newly written derivatives contracts to reference a riskless rate, rather than LIBOR (or LIBOR+). For example, bank regulators could, for the firms that fall under their authority, increase the effective capital charges that apply to derivatives based on LIBOR relative to those based on a riskless rate. In addition to mitigating manipulation incentives, we believe that there is a legitimate safety-and-soundness rationale for doing so. As noted above, the survey-based nature of current LIBOR creates the risk of defections from the bank reporting panels, with the attendant dangers of market-wide disruptions if the LIBOR rate cannot be produced. To the extent that a transition from LIBOR to LIBOR+ takes a long time or is subject to uncertainty, beginning the process of moving derivatives to an alternative reference rate would have the added benefit of reducing this type of risk to financial stability.

It is difficult to say just how much regulatory pressure would need to be applied to substantially change contracting practices in the derivatives market, or how much force it would be appropriate to apply. On the one hand, as we have argued above, there are elements of a pure coordination problem here. It may be that many derivatives users would actually prefer to be in an equilibrium in which there were highly liquid contracts that referenced a riskless rate, as opposed to an equilibrium in which the benchmark contains a significant spread component for bank credit risk. In this case, a strong regulatory hand that pushes the outcome towards this new equilibrium would be seen as socially desirable. On the other hand, there is undoubtedly significant heterogeneity among derivatives users, and it is far from clear that all would prefer the new equilibrium. As a result, any strong push by regulators would likely create losers as well as winners, which cuts against an overly aggressive use of regulatory authority such as a highly punitive capital charge on derivatives that remain linked to LIBOR or LIBOR+. Striking the right balance on this dimension seems to us to be one of the most challenging aspects of the reform process.

Conclusion

Rather than restating our arguments, we close by highlighting a fundamental limitation of our analytical approach. From the outset, we have taken as given two policy objectives: 1) that it is desirable to maintain large, deep, and liquid interest-rate derivatives markets; and 2) that it is also desirable to design markets in a way that leans against manipulation. But as we have seen, there is a fundamental tension between these two objectives: the deeper and more liquid a derivatives market becomes, the more tempting it is for market participants to manipulate the underlying benchmark referenced by these derivatives.

This suggests that even the best market design can only go so far, and that if one wishes to support the existence of a very large derivatives market, some equilibrium level of manipulation may be an inevitable cost of doing business. This is an uncomfortable prospect for policymakers to acknowledge, but it is nevertheless important to be candid on this score. The last thing one wants is to embark on a costly and time-consuming set of reforms while overpromising what they can deliver. On a more constructive note, acknowledging the limits of market-design policies, such as those suggested here, underscores the need for a complementary attack on the manipulation problem from a legal (rules and enforcement) angle. Given that one cannot count on market design as a panacea for preventing manipulation, vigorous enforcement of the civil and criminal statutes against market manipulation will continue to play an important role no matter what other reforms are undertaken.

References

Abrantes-Metz, Rosa, Michael Kraten, Albert Metz, and Gim S. Seow. 2012. "Libor Manipulation?" *Journal of Banking and Finance* 36(1): 136–50.

Bank for International Settlements. 2013. "Triennial Central Bank Survey: OTC Interest Rate Derivatives Turnover in April 2013: Preliminary Global Results." Monetary and Economic Department, Bank for International Settlements, September. http://www.bis.org/publ/rpfx13ir.pdf.

Brundsen, Jim. 2014. "ECB Sees Rules to Stem Bank Exodus from Benchmark Panels." *Bloomberg,* June 19. http://www.bloomberg.com/news/2014-06-19/ecb-seeks-rules-to-stem-bank-exodus-from-benchmark-panels.html.

Duffie, Darrell, and Piotr Dworczak. 2014. "Robust Benchmark Design." http://www.darrellduffie.com/uploads/working/DuffieDworczakJune2014.pdf.

Duffie, Darrell, Piotr Dworczak, and Haoxiang Zhu. 2014. "Benchmarks in Search Markets." http://ssrn.com/abstract=2515582.

Duffie, Darrell, David Skeie, and James Vickery. 2013. "A Sampling-Window Approach to Transactions-Based Libor Fixing." Federal Reserve Bank of New York Staff Report 596, February. http://www.newyorkfed.org/research/staff_reports/sr596.pdf.

Fleming, Michael, John Jackson, Ada Li, Asani Sarkar, and Patricia Zobel. 2012. "An Analysis of OTC Interest Rate Derivatives Transactions: Implications for Public Reporting." Federal Reserve Bank of New York Staff Report 557, March 2012, revised October 2012. http://www.newyorkfed.org/research/staff_reports/sr557.pdf.

Foreign Exchange Benchmark Group. 2014. "Foreign Exchange Benchmarks." Consultative Document. Financial Stability Board, July 15. http://www.financialstabilityboard.org/publications/r_140715.pdf.

Gorton, Gary B., Andrew Metrick, and Lei Xie. 2014. "The Flight from Maturity." NBER Working Paper 20027, April. http://www.nber.org/papers/w20027.pdf.

Hou, David, and David Skeie. 2013. "LIBOR: Origins, Economics, Crisis, Scandal and Reform." *The New Palgrave Dictionary of Economics, Online Edition,* 2013, edited by Steven N. Durlauf and Lawrence E. Blume. http://www.dictionaryofeconomics.com/article?id=pde2013_L000246.

International Organization of Securities Commissions. 2013. "Principles for Financial Benchmarks: Final Report." FR07/13. July. http://www.iosco.org/library/pubdocs/pdf/IOSCOPD415.pdf.

Kuo, Dennis, David Skeie, and James Vickery. 2012. "A Comparison of Libor to other Measures of Banks' Borrowing Costs." June. http://www.newyorkfed.org/research/economists/vickery/LiborKSV_staff_webpage.pdf.

Market Participants Group on Reforming Interest Rate Benchmarks. 2014. *Final Report.* Financial Stability Board, March, 14. http://www.financialstabilityboard.org/publications/r_140722b.pdf.

McCauley, Robert N. 2001. "Benchmark Tipping in the Money and Bond Markets." A section in the *BIS Quarterly Review,* March 2001, pp. 39–45.

Mollenkamp, Carrick. 2008. "Bankers Cast Doubt on Key Rate Amid Crisis." *Wall Street Journal,* April 16.

Mollenkamp, Carrick, and Mark Whitehouse. 2008. "Study Casts Doubt on Key Rate." *Wall Street Journal,* May 29.

Official Sector Steering Group. 2014. "Reforming Major Interest Rate Benchmarks." Financial Stability Board, July 22. http://www.financialstabilityboard.org/publications/r_140722.pdf.

Ridley, Kristin, and Huw Jones. 2012. "Insight: A Greek Banker, the Shah and the Birth of Libor." Reuters, August 7. http://www.reuters.com/article/2012/08/08/us-banking-libor-change-idUSBRE87702320120808.

Snider, Connan, and Thomas Youle. 2012. "The Fix is In: Detecting Portfolio Driven Manipulation of the LIBOR." September. http://www.dartmouth.edu/~tyoule/documents/main_paper_2(1).pdf.

United Kingdom Financial Conduct Authority. 2012. *The Wheatley Review of LIBOR: Final Report.* September. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/191762/wheatley_review_libor_finalreport_280912.pdf.

US Commodities Futures Trading Commission. 2012. United States of America before the Commodities Futures Trading Commission, in the matter of Barclays PLC, Barclays Bank PLC and Barclays Capital Inc. CFTC Docket No. 12-25. http://www.cftc.gov/ucm/groups/public/@lrenforcementactions/documents/legalpleading/enfbarclaysorder062712.pdf.

Vaughn, Liam. 2014. "Gold Fix Study Shows Signs of Decade of Bank Manipulation." *Bloomberg.* February 28. http://www.bloomberg.com/news/2014-02-28/gold-fix-study-shows-signs-of-decade-of-bank-manipulation.html.

Bitcoin: Economics, Technology, and Governance[†]

Rainer Böhme, Nicolas Christin, Benjamin Edelman, and Tyler Moore

Bitcoin is an online communication protocol that facilitates the use of a virtual currency, including electronic payments. Since its inception in 2009 by an anonymous group of developers (Nakamoto 2008), Bitcoin has served approximately 62.5 million transactions between 109 million accounts. As of March 2015, the daily transaction volume was approximately 200,000 bitcoins—roughly $50 million at market exchange rates—and the total market value of all bitcoins in circulation was $3.5 billion (Blockchain.info 2015). Table 1 summarizes Bitcoin activity to date. (We will follow the convention in the computer science literature of using capital-B Bitcoin to refer to the system, and lower-b bitcoin to refer to the unit of account.)

Bitcoin's rules were designed by engineers with no apparent influence from lawyers or regulators. Rather than store transactions on any single server or set of servers, Bitcoin is built on a transaction log that is distributed across a network of participating computers. It includes mechanisms to reward honest participation, to bootstrap acceptance by early adopters, and to guard against concentrations of power. Bitcoin's design allows for irreversible transactions, a prescribed path of money creation over time, and a public transaction history. Anyone can create a

■ *Rainer Böhme is Professor of Security and Privacy, University of Innsbruck, Innsbruck, Austria. Nicolas Christin is Assistant Research Professor, Department of Electrical and Computer Engineering and CyLab, Carnegie Mellon University, Pittsburgh, Pennsylvania. Benjamin Edelman is Associate Professor of Business Administration, Harvard Business School, Boston, Massachusetts. Tyler Moore is Assistant Professor of Computer Science and Engineering, Southern Methodist University, Dallas, Texas. Their email addresses are rainer.boehme@uibk.ac.at, nicolasc@cmu.edu, bedelman@hbs.edu, and tylerm@smu.edu.*

[†] To access the disclosure statements, visit
http://dx.doi.org/10.1257/jep.29.2.213

doi=10.1257/jep.29.2.213

Table 1
Bitcoin Activity to Date
(as of March 2015)

Total bitcoins minted	≈ 14 million
US dollar equivalent at market price	≈ 3.5 billion
Total number of reachable Bitcoin nodes	≈ 6,500[a]
Total (cumulative) number of transactions	≈ 62.5 million
Total number of accounts ever used	≈ 109 million
Block chain size	≈ 30.3 GB
Number of blocks to date	≈ 350,000
Estimated daily transaction volume	≈ 200,000 BTC (≈ $50 million)
Average transaction value	≈ 2 BTC (≈ $500)[b]
Computation invested in puzzle solutions	≈ 4,254 exaflops[c]
Power consumption	> 173 MW (continuously)[d]

Source: Authors' compilation and own computations derived from (Yeow 2015; Blockchain.info, 2015; Bitcoincharts.com, 2015; Bitcoin Wiki 2015b).

[a] Reports only publicly reachable notes and excludes "private" nodes, for example, nodes hosted on private networks behind a firewall, which are likely to represent the majority of the network but cannot be reliably measured.

[b] Excludes change. The distribution is skewed toward small transactions. We estimate the median transaction amount to be around 0.02 bitcoins ($5).

[c] This corresponds to roughly 11,500 times the combined power of the top 500 supercomputers in the world. That said, supercomputers can perform all sorts of mathematical operations, while Bitcoin miners are generally highly specialized in a single type of cryptographic operation.

[d] Reflects a computation similar to Bonneau's (2014) lower bound. According to Bitcoin Wiki (2015b), the most energy-efficient mining hardware can perform 1,957 millions of cryptographic operations ("hashes") per Joule (W/s). The current aggregate power of the Bitcoin network is 340,000 terahashes (10^{12}) per second (Bitcoincharts.com 2015). This capacity would require continuous consumption of 173 MW, if every miner used the most energy-efficient hardware.

Bitcoin account, without charge and without any centralized vetting procedure—or even a requirement to provide a real name. Collectively, these rules yield a system that is understood to be more flexible, more private, and less amenable to regulatory oversight than other forms of payment—though as we discuss in subsequent sections, all these benefits face important limits.

Bitcoin is of interest to economists as a virtual currency with potential to disrupt existing payment systems and perhaps even monetary systems. Even at their current early stage, such virtual currencies provide a variety of insights about market design and the behavior of buyers and sellers. This article presents the platform's design principles and properties for a nontechnical audience; reviews its past, present, and future uses; and points out risks and regulatory issues as Bitcoin interacts with the conventional financial system and the real economy.

Bitcoin Design Principles

Scarcity is a prerequisite for ascribing value to any form of money. At a micro level, scarcity protects against counterfeiting. More broadly, scarcity bounds the growth path of the monetary base and facilitates price stability. In modern economies, where money is held in electronic forms, scarcity is preserved by legal rules ensuring the correctness of bookkeeping records: that is, electronic money involves a financial system in which transactions trigger a credit for one account and a corresponding debit to another. Central banks hold the power to adjust the absolute quantity of money in circulation.

Against this backdrop, Bitcoin can be understood as the first widely adopted mechanism to provide absolute scarcity of a money supply. By design, Bitcoin lacks a centralized authority to distribute coins or to track who holds which coins. Consequently, the process of issuing currency and verifying transactions is considerably more difficult than in classic bookkeeping systems. Meanwhile, Bitcoin issues new currency to private parties at a controlled pace in order to provide an incentive for those parties to maintain its bookkeeping system, including verifying the validity of transactions.

Enabling Technologies and Processes

The "Bitcoin core" software can be freely downloaded at https://bitcoin.org/en/choose-your-wallet. The standard Bitcoin implementation includes a number of features. Typically, it creates a "wallet" file for the user that can store bitcoins (without giving a name or proof of identity); it creates an individual node for the user in the peer-to-peer Bitcoin network that can be used with a standard Internet connection; and it provides access to the "block chain" data structure that verifies all past Bitcoin activity.

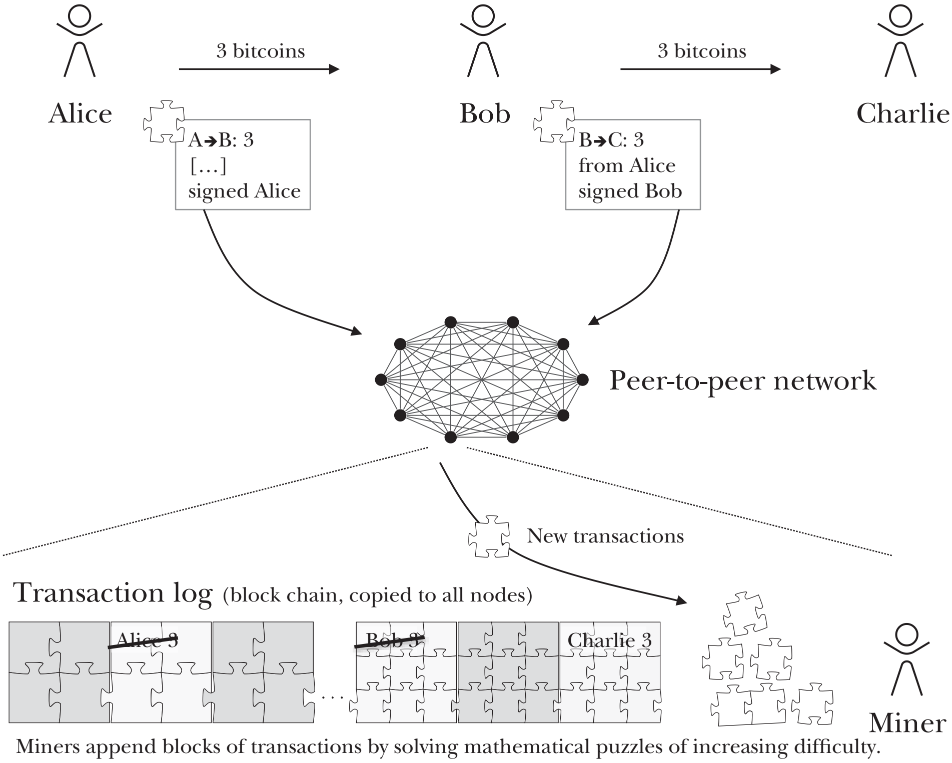
Transactions and the Block Chain

Bitcoins are recorded as transactions. For instance, some user Charlie does not simply "hold" three bitcoins. Rather, Charlie participates in a publicly verifiable transaction showing that he received three bitcoins from Bob. Charlie was able to verify that Bob could make that payment because there was a prior transaction in which Bob received three bitcoins from Alice and there was no prior transaction in which Bob spent these three bitcoins. Figure 1 illustrates these interactions.

Indeed, each individual bitcoin can readily be traced back through all transactions in which it was used, and thus to the start of its circulation. All Bitcoin transactions are readable by everyone in records stored in a widely replicated data structure. In general, transactions are ordered recursively by having the input of a transaction (roughly, the source of funds) refer to the output of a previous transaction. (For example, the transaction might reveal that Bob pays Charlie using bitcoin he received from Alice.)

Figure 1

Bitcoin's Approach to Transaction Flow and Validation



Source: Authors.

Bitcoin relies on two fundamental technologies from cryptography: public-private key cryptography to store and spend money; and cryptographic validation of transactions. Standard public-private key cryptography lets anyone create a public key and an associated private key (Diffie and Hellman 1976). Public keys are designed to be widely shared—hence the name. Messages encrypted with a public key can only be descrambled by someone who possesses the corresponding private key, allowing anyone to encrypt a message that only the specified recipient can read. Similarly, messages encrypted with a private key can only be descrambled with the corresponding public key, allowing a specified sender to create a message that can be confirmed to be authentic. Public-private key cryptography is widely used: in the best-known example, web browsers on a HTTPS "secure website" encrypt communications with that site's advertised public key in order to begin a secure connection. In Bitcoin, similar encryption fundamentals authenticate instructions to transfer money to other participants. Such an instruction is encrypted using the sender's private key, confirming for everyone that the instruction in fact came from the sender.

Suppose that Alice has three bitcoins that she wants to give to Bob. She publishes a message in the Bitcoin network indicating that she is transferring three of her existing bitcoins, along with a reference to the transaction where she had received those bitcoins. Part of this message is encrypted by Alice's private key to prove that the instruction came from her, in a method akin to a signature on a paper check. Later, if Bob wants to send bitcoins to Charlie, he publishes a message, again encrypted with his private key, indicating that he got his bitcoins from Alice and what he wants to send to whom. The Bitcoin network identifies Alice, Bob, and Charlie only by their public keys, which serve as account numbers.

Every new transaction that is published to the Bitcoin network is periodically grouped together in a "block" of recent transactions. To make sure no unauthorized transactions have been inserted, the block itself is compared to the most recently published block—yielding a linked sequence of blocks, or "block chain." A new block is added to the chain roughly every ten minutes. With this data structure in place, any Bitcoin user can verify that a prior transaction did in fact occur.

Keeping the transaction record operational and updated is a public good, as it is the foundation of the entire Bitcoin system. To encourage users to assist, the Bitcoin system periodically awards newly minted bitcoins to the user who solves a mathematical puzzle that is based on the pre-existing contents of the block (which prevents tampering with the block and hence modifying prior transactions) and which can only be solved by computationally intensive methods that include a random component. Thus, faster computing is more likely to solve a given problem and will solve a greater number of these problems, but speed alone will not guarantee success.

Upon solving the puzzle, the user publishes a "block" which contains a proof-of-work that a solution was carried out along with all observed transactions that have taken place since the last puzzle solution was announced and a reference to the previous complete block. After other users verify the solution, they start working on a new block containing new outstanding transactions. This process is called "mining" and recursively ensures that the total historical ordering on all blocks ("chain") is agreed by the entire network.

A Bitcoin transaction does not clear (and hence is not final) until it has been added to the consensus block chain. Transaction batches are added every ten minutes on average. However, miners are continuously working on adding blocks of transactions, and building on previous transactions. By continually presenting their solutions to the puzzles, with the associated new tail of the block chain, miners are in effect "voting" on the correct record of Bitcoin transactions, and in that way verifying the transactions. In some cases, a transaction batch will be added to the block chain, but then a few minutes later it will be altered because a majority of miners reached a different solution. Sources typically recommend considering a Bitcoin transaction final only after six confirmations, to assure that the transaction is truly recorded in a permanent part of the block chain. While this provides greater assurance, it creates a delay of approximately one hour before a Bitcoin transaction is finally validated.

As miners update the block chain, their computational efforts carry significant costs. In particular, the computerized proof-of-work calculations are quite power-intensive, consuming more than 173 megawatts of electricity continuously. For perspective, that amount is approximately 20 percent of an average nuclear power plant (World Nuclear Association 2015), or approximately $178 million per year at average US residential electricity prices. These computational costs have grown sharply and may rise further because Bitcoin automatically adjusts puzzle difficulty so that the time interval between two blocks remains roughly ten minutes. As more computing power joins the Bitcoin system, the puzzles automatically become more difficult, increasing computing and electricity requirements. In fact, an arms race ensued as the price of bitcoin rose. Taylor (2013) compares the difficulty of solving the puzzle to the bitcoin-dollar exchange rate, finding that spikes in the exchange rate—bitcoins becoming more valuable in terms of US dollars—have been followed by increases in computational difficulty.

Built-in Incentives

Bitcoin includes several built-in incentives to encourage useful behavior. The miners who verify the block chain are rewarded with—what else?—bitcoins. At first, miners solving the puzzle received a reward of 50 bitcoins. This reward is periodically cut in half, and it stands at 25 as of March 2015. After 21 million bitcoins have been minted, the reward falls to zero and no further bitcoins will be created. Hence, the protocol design for Bitcoin sets a controlled pace for the expansion of the currency and an ultimate limit to the number of bitcoins issued.

Miners have a second potential source of revenue (which will become the only source of revenue once all bitcoins have been created). When listing a transaction, the buyer and seller can also offer to pay a "transaction fee," which is a bonus payment to whatever miner solves the puzzle that verifies the transaction. These fees are optional, but 97 percent of the transactions in 2014 include a fee, most often set at the default rate of the standard client software, 0.0001 bitcoin. In relative terms, the transaction fees are below 0.1 percent of total transaction value (Möser and Böhme 2014). However, as the mathematical puzzles become harder, there will presumably be a point where the automatic reward for solving the puzzle drops below the cost of doing so. At that point, one possibility is that those who wanted a Bitcoin transaction could bid up the optional fees. Houy (2014a) models equilibria for the level of transaction when the minting reward drops below the cost of mining.

Early in Bitcoin's operation, updating the block chain yielded bitcoins more often and hence more readily per unit of computing power provided. This design benefited those who ran the Bitcoin platform at the outset—helping to create the critical mass needed to bootstrap the platform (Böhme 2013). Today, some users still find mining profitable, but effective mining now requires specialized hardware (particularly well-suited to solving the mathematical puzzles at issue) as well as access to low-cost electricity.

Requiring miners to solve a puzzle helps avoid certain types of fraud. In principle, a system like Bitcoin could validate transactions using a simple consensus by majority vote, with a majority of connected users able to affirm that a given transaction in fact occurred. But then an attacker could game the system by creating numerous fake identities. In response, the Bitcoin protocol makes it costly to submit fake votes. Consistent with the Internet's open architecture, anyone can connect multiple computers to the Bitcoin system. But voting on the authenticity of a transaction requires first working to solve a mathematical puzzle that is computationally hard to solve (although easy to verify). Solving the puzzle provides "proof of work"; in lieu of "one person, one vote," Bitcoin thus implements the principle of "one computational cycle, one vote." Through this design, the proof-of-work mechanism simultaneously discourages creating numerous fake identities and also provides incentives to participate in verifying the block chain.

What Bitcoin Doesn't Have

Compared with conventional payment systems, Bitcoin lacks a governance structure other than its underlying software. This has several implications for the functioning of the system. First, Bitcoin imposes no obligation for a financial institution, payment processor, or other intermediary to verify a user's identity or cross-check with watch-lists or embargoed countries. Second, Bitcoin imposes no prohibition on sales of particular items; in contrast, for example, credit card networks typically disallow all manner of transactions unlawful in the place of sale (MacCarthy 2010). Finally, Bitcoin payments are irreversible in that the protocol provides no way for a payer to reverse an accidental or unwanted purchase, whereas other payment platforms, such as credit cards, do include such procedures. As discussed in subsequent sections, these design decisions are intentional—simplifying the Bitcoin platform and reducing the need for central arbiters, albeit raising concerns for some users.

Centralization and Decentralization in the Bitcoin Ecosystem

The key innovation in Bitcoin, compared to other forms of cryptographic cash (Chaum 1983) or virtual currencies (European Central Bank 2012), is its decentralized core technologies. Early adopters praised decentralization and by all indications chose Bitcoin because they wanted to use a decentralized system (Raskin 2013). Decentralization offers certain advantages. It avoids concentrations of power that could let a single person or organization take control. It often promotes availability and resiliency of a computer system, avoiding a central point of failure. It offers at least the appearance of greater privacy for users (and perhaps greater genuine privacy) because in theory an eavesdropping adversary cannot observe transactions across the system by targeting any single point or any single server. (However, as we discuss below, significant privacy concerns remain.)

Nonetheless, the decentralization touted by Bitcoin has not fully come to fruition. While the Bitcoin protocol supports complete decentralization (including the possibility of all participants acting as miners), significant economic forces push towards de facto centralization and concentration among a small number of intermediaries at various levels of the Bitcoin ecosystem. We review four key categories of intermediaries that have shaped Bitcoin's evolution: currency exchanges, digital wallet services, mixers, and mining pools. A fifth type of intermediary, payment processors, is discussed further below.

Currency Exchanges

Currency exchanges allow users to trade bitcoins for traditional currencies or other virtual currencies. Most operate double auctions with bids and asks much like traditional financial markets, and charge a commission ranging from 0.2 to 2 percent. Some exchanges offer more advanced trading tools, such as limit or stop orders. To date, derivatives markets and short-selling remain rare.

At present, many trades in bitcoin are accompanied by one or even two conversions from and/or to conventional currencies. Furthermore, price quotes in bitcoin are almost always computed in real time by reference to a fixed amount of conventional currency. Thus, Bitcoin today resembles more a payment platform than what economists consider a currency.

While few technical barriers impede setting up intermediaries in the Bitcoin ecosystem, there are significant regulatory requirements. In the United States, currency exchanges generally operate as "money transmitters" and thus must register with the Financial Crimes Enforcement Network (FinCEN) as money services businesses. Registration includes a state-by-state licensing requiring both legal fees and posting bonds. Certification in a single state often costs at least $10,000, so nationwide participation can easily reach six figures on fees alone. Other countries have broadly similar rules. In Germany, currency exchanges that manage deposits on behalf of clients are viewed as "deposit banks" with a minimum capital requirement of €5 million.

In addition, currency exchanges need online infrastructure capable of withstanding attacks including hacking and denial-of-service attacks. For these reasons, the number of Bitcoin exchanges has remained modest, and the number of Bitcoin exchanges with significant volume has been even smaller. In spring 2012, the Japan-based Mt. Gox exchange served over 80 percent of all Bitcoin transactions. However, Mt. Gox collapsed in early 2014 and reported in its bankruptcy filing "losing" 754,000 of its customers' bitcoins worth approximately $450 million at the time of closure (Abrams, Matthew, and Tabuchi 2014). In March 2015, the seven largest exchanges were BTC China, OKCoin, Huobi, Bitfinex, LakeBTC, Bitstamp, and BTC-e, which jointly served more than 95 percent of all bitcoin trade from October 2014 to March 2015 (Bitcoinity.org 2015).

Digital Wallet Services

Bitcoin wallets are data files that include Bitcoin accounts, recorded transactions, and private keys necessary to spend or transfer the stored value. Some users install specialized wallet software (such as Armory, Electrum, or Hive) on their personal devices to maintain control over their bitcoins. However, many users find this task unappealing. Bitcoin wallet software can be difficult to install, and can impose onerous technical requirements—such as storing a copy of the entire block chain, which was 30 gigabytes as of March 2015. (Not all participants need to download the entire chain, but the system does rely on some users electing to do so.) Other users worry about security: a crash or attack on the computer holding the digital wallet could cause the loss of a user's bitcoins.

As a result, many users rely on a digital wallet service that keeps the required files on a shared server with access via the web or via phone-based apps. A key distinction among digital wallet services is whether the service knows the account's private key. Some services (including Blockchain.info, StrongCoin, and CoinPunk) let the user maintain control over private keys, meaning that the service is incapable of spending the user's bitcoin (nor could hackers do so even if they fully infiltrated the wallet service). For such firms, the user must keep and present the private key when needed, and a user who loses the key or allows it to be compromised is at high risk. In contrast, other services (such as Coinbase and Xapo) require users to let the service store their private keys, which increases risk if the digital wallet service is compromised. In practice, digital wallet services tend to increase centralization—either expanding the role and importance of exchanges, or adding an additional service that is likely to be centralized due to high fixed costs, low marginal costs, and limited diversity in users' needs.

Mixers

As initially envisioned, the Bitcoin transaction log shows each transaction made from each payer to each payee, along with the public keys serving as pseudonyms of each. As a result, anyone who knows the identity of any user from any transaction—perhaps the mailing address used for delivery of purchased goods, or the bank account used to purchase bitcoins—can track that user's other transactions made with the same pseudonym, both before and since.

To preserve privacy against this tactic, *mixers* let users pool sets of transactions in unpredictable combinations, thus preventing tracking across transactions. Suppose Alice wants to pay Bob one bitcoin, and Charles wants to pay Daisy one bitcoin. To mislead an observer who tracks these payments, Alice and Charles could both pay a mixer "Minnie" and provide additional confidential instructions for Minnie to pay Bob and Daisy one bitcoin each. An observer would see flows from Alice and Charles to Minnie, and from Minnie to Bob and Daisy, but would not be able to tell whether it was Alice or Charlie who sent money to Bob. In practice, mixers must ensure that timing does not yield clues about money flows, which is particularly difficult since it is rare for different users to seek to transmit the exact same amount. Mixers have been used to promote anonymity in online communications, most famously by the Tor network, so their limitations are now widely known (Danezis and Diaz 2008). In addition to standalone services, some mixers are incorporated as a feature provided by digital wallets.

While mixers seem to improve privacy, they create additional challenges. For one, the finality of Bitcoin payments leaves payers with little recourse if a mixer absconds with their funds. Furthermore, mixing protocols are usually not public, so their effectiveness cannot be proven. Indeed, correlations in timing might still reveal transaction counterparts, particularly at little-used mixers (Möser, Böhme, and Breuker 2013). Finally, mixers charge 1 to 3 percent of the amount sent, increasing costs for those who choose to use them.

Mining Pools

As discussed above, bitcoins are created when a miner successfully solves a mathematical puzzle. The puzzles have become significantly more difficult over time, and lumpy rewards mean a lone miner is now at risk of contributing resources in an attempt to solve a puzzle but then receiving no reward. In response, mining pools now combine resources from numerous miners. Miners work independently, but upon winning a miner shares earnings with others in the pool (much like consumers sharing resources to buy lottery tickets). As of March 2015, the two largest pools are AntPool and F2Pool, which together account for around one-third of Bitcoin mining activities.

Oversized mining pools threaten the decentralization that underpins Bitcoin's trustworthiness. In several instances including a twelve-hour interval in June 2014, GHash briefly held more than 50 percent of total mining power, which could have allowed GHash pool operators to attempt manipulations. An attacker who holds a majority of Bitcoin's computational resources can alter some of the system's records, including inserting false transactions and rejecting actual transactions (albeit with a strong chance that others will notice), or deviate from the protocol rules.

Uses of Bitcoin

Early: Silk Road and Other Illicit Activities

After early proof-of-concept transactions, the first notable adopters of Bitcoin were businesses that sought features not easily available through alternatives: greater anonymity and the absence of rules concerning what could be bought or sold.

One prominent example involved the online sale of narcotics including marijuana, prescription drugs, and benzodiazepines (a class of psychoactive drugs). Drugs had been sold online for years, typically on informal bulletin boards and on websites such as "The Farmer's Market," a website that listed various narcotics available for purchase with payment using other services including PayPal (Kim 2014). When Bitcoin is used with tools to anonymize network traffic such as Tor (Dingledine, Mathewson, and Syverson 2004), marketplaces could provide stronger assurances of anonymity. Transaction volume grew sharply: Christin (2013) estimates that the turnover on the Silk Road anonymous online marketplace, the first to support Bitcoin transactions exclusively, reached $15 million per year just one year after it began operation. Silk Road's own category classifications confirm the prevalence of narcotics items, which dominated Silk Road's top categories as shown in Table 2. Examining 30 months of Silk Road data from February 2011 to July 2013,

Table 2
The Ten Most Popular Product Categories on the Silk Road Website in January–July 2012

Category	*Number of items*	*Percentage*
Weed	3,338	13.7%
Drugs	2,193	9.0%
Prescription	1,784	7.3%
Benzodiazepines	1,193	4.9%
Books	955	3.9%
Cannabis	877	3.6%
Hash	820	3.4%
Cocaine	630	2.6%
Pills	473	1.9%

Source: Christin (2013).
Note: Categories are self-reported by sellers.

the government evidence in the case against Ross Ulbricht lists 9.9 million bitcoins of transactions, which, accounting for the varying exchange rates, corresponds to $214 million (US v. Ulbricht, 2014, Government Exhibit 940). After the demise of Silk Road at the hands of law enforcement (discussed further below), alternative markets opened in its stead—a "new" Silk Road, as well as more than 30 competitors—and it is unclear whether the Silk Road takedown actually reduced contraband activity using Bitcoin.

While litigation documents largely focus on Silk Road as a marketplace for drugs and other contraband, the site's general-purpose platform stood ready to sell *anything*. Reputation systems ensured trustworthiness of the transaction parties; escrow services mitigated counterparty risk; and, in some cases, hedges protected customers against currency volatility. Criminal charges criticized Silk Road's fees: for escrow service, these averaged 8 percent in comparison to credit card system fees of approximately 3 percent—allegedly an indicator of Silk Road's distinctive profit from misbehavior. But eBay's fees typically somewhat exceed Silk Road's fees, calling into question whether high fees in and of themselves indicate a platform's purpose or responsibility.

Silk Road sellers appear to have exploited some arbitrage opportunities. For instance, marijuana is generally cheaper in the Netherlands than in Australia, providing Netherlands-based Silk Road sellers an opportunity to compete advantageously with street sellers in Australia. Numerous online discussions flagged this opportunity and the sellers who invoked it, and analysis of Silk Road's transactions confirms disproportionate items sold from the Netherlands.

Gambling sites also turned to Bitcoin, both to protect customer privacy and to receive funds from customers unable to use other payment methods. The most popular single Bitcoin gambling game is Satoshi Dice, a simple betting game in which a player wins if a dice roll is less than the player's chosen number. This service reported 2012 earnings of approximately 33,000 bitcoins (or roughly $403,000 at then-applicable rates) with an average monthly growth of 78 percent at the time (Matonis 2013). For several months, the service's (low value) payments accounted for up to 80 percent of total Bitcoin transactions (Möser and Böhme 2014). The Bitcoin Wiki (2015a) now reports around 100 casinos, poker sites, dice games, lotteries, and betting services.

Bitcoin can also be used to evade international capital controls. In December 2013, the People's Bank of China, the central bank of China, banned Chinese banks from relationships with Bitcoin exchanges, a decision which the *Economist* magazine attributed to a desire to prevent yuan from being moved overseas via Bitcoin (D.K. 2013). Similarly, interest in Bitcoin appears to be particularly high in Argentina, where government policy strictly limits transfers to other currencies (McLeod 2013).

Current: Consumer Payments, Buy-and-Hold

In light of widespread criticism of the fees charged by credit and debit card networks (Anderson 2012), Bitcoin could offer an alternative that might pressure card networks to lower their prices to merchants. Some early evidence seems to confirm that Bitcoin may have this effect. Overstock.com, an online retailer, began to receive payments by Bitcoin in January 2014. Overstock reported a favorable response, including significant revenue gains, large average order sizes, and desirable customer demographics (Sidel 2014). Other merchants subsequently added Bitcoin support, including Expedia (travel), Newegg (electronics), Foodler (restaurant delivery and takeout), Gyft (gift cards for dozens of merchants), and TigerDirect (electronics). Payment processors help online merchants adjust their websites to accept Bitcoin. Early user reviews are mixed: users seem largely satisfied, though technical glitches sometimes occur. Merchants appear particularly pleased because Bitcoin payment processing is strikingly low-cost for them. For example, Coinbase (a payment processing firm) currently charges zero percent on incoming payments up to $1 million per merchant per annum, and 1 percent thereafter, which is considerably lower than the fees that merchants bear when a credit card is used to pay for a purchase.

It is less clear that consumers benefit from paying by Bitcoin. Many credit cards provide consumers with rebates of 1 percent, 2 percent or even more, as well as benefits of similar value such as frequent flyer points and merchandise credits. A consumer who pays by Bitcoin loses such rebates or bonuses. Edelman (2014) points out that even if a consumer already has bitcoins, the consumer would be better off making a purchase with a 1.5 percent cashback credit card, paying a 1 percent fee to convert bitcoins to dollars, then using those dollars to pay the credit card bill. Some merchants have responded by providing additional benefits to consumers who pay by Bitcoin: for example, Overstock provides a 1 percent rebate. However, if competing Bitcoin exchanges bid the 1 percent fee for converting from currency to bitcoin downwards, there could be room to make both consumers and merchants better off than through payments by credit card.

The block chain poses a further barrier to using Bitcoin for general-purpose payments. Every Bitcoin transaction, large or small, must be copied into all future versions of the block chain. If Bitcoin expanded to include a huge volume of transactions—as from millions of users' small day-to-day payments—the storage burden would need to be addressed. Furthermore, updating the block chain entails an undesirable delay, making Bitcoin too slow for many in-person retail payments.

Meanwhile, other users appear to be buying bitcoins not to use them but to hold them in appreciation. Meiklejohn, Pomarole, Jordan, Levchenko, McCoy, Voelker, and Savage (2013) finds that of the bitcoins mined in 2009–2010, more than 60 percent remain unspent or took more than one year to be spent.

Overall, some question whether the growth of Bitcoin payments is actually as rapid as one might expect for a successful payments service. Evans (2014) compares Bitcoin's growth to that of mPesa, a widely used person-to-person payment system using mobile phones in Kenya. Aligning the services based on months since launch, Evans finds Bitcoin's adoption less than one-twentieth as rapid.

Possible and Future: General-Purpose Payments, Mainstream Store of Value, and Enabling Technology

Some proponents envision Bitcoin evolving into an all-purpose payment mechanism. If a payer already held bitcoins and if a payee was content to retain bitcoins rather than convert to a traditional currency, fees would be relatively low: the only costs are transaction fees paid to the successful miner who solved that block's puzzle (and perhaps also a small minting reward). However, to date most payments entail at least one party needing to convert to or from bitcoin, which adds to transaction costs. Overstock.com, the first prominent retailer to accept bitcoins, reports keeping 10 percent of its bitcoin gross receipts in that form (Sidel 2014), but given Overstock's net margin of 0.6 percent (per its 2014 SEC 10-K), this effectively requires transferring profits from the company's other operations.

It might seem natural for consumers to use Bitcoin for international remittances, which may sometimes cost $50 or more, rather than as a substitute for credit card payments where consumers often receive a rebate. But so far, there is little sign of Bitcoin use in this area. The fees from services such as Western Union may appear high at first glance. But Western Union also offers a suite of services including accepting and dispensing cash, which is distinctively useful in low-income countries where transfer from bitcoin to local currency is likely to be difficult and where merchants are unlikely to accept payment by Bitcoin.

Some computer scientists and entrepreneurs report excitement at Bitcoin not for its role in facilitating payments, but for its ability to create a decentralized record of almost anything. Marc Andreessen (2014), best known as coauthor of Mosaic (the first widely-used web browser), presented the rationale:

> Bitcoin gives us, for the first time, a way for one Internet user to transfer a unique piece of digital property to another Internet user, such that the transfer is guaranteed to be safe and secure, everyone knows that the transfer has taken place, and nobody can challenge the legitimacy of the transfer. . . . All these are exchanged through a distributed network of trust that does not

> require or rely upon a central intermediary like a bank or broker. What kinds of digital property might be transferred in this way? Think about digital signatures, digital contracts, digital keys (to physical locks, or to online lockers), digital ownership of physical assets such as cars and houses, digital stocks and bonds . . . and digital money.

To date, there has been only limited use of the Bitcoin platform to provide services other than payment. Entrants building on the Bitcoin platform include Namecoin, an alternative domain name system; Colored Coins, a means to manage virtual property rights (Rosenfeld 2012); CommitCoin, a secure commitment scheme (Clark and Essex 2012), a timed version of which can be repurposed to ensure fairness in multi-party computation (Andrychowicz, Dziembowski, Malinowski, and Mazurek 2014) in order to run auctions without an auctioneer; and FutureCoin (Clark, Bonneau, Felton, Kroll, Miller, and Narayanan 2014), which enables decentralized prediction markets. However, none of these startups has attracted large-scale use to date, and each faces significant competition from firms and processes using more traditional system design.

Risks in Bitcoin

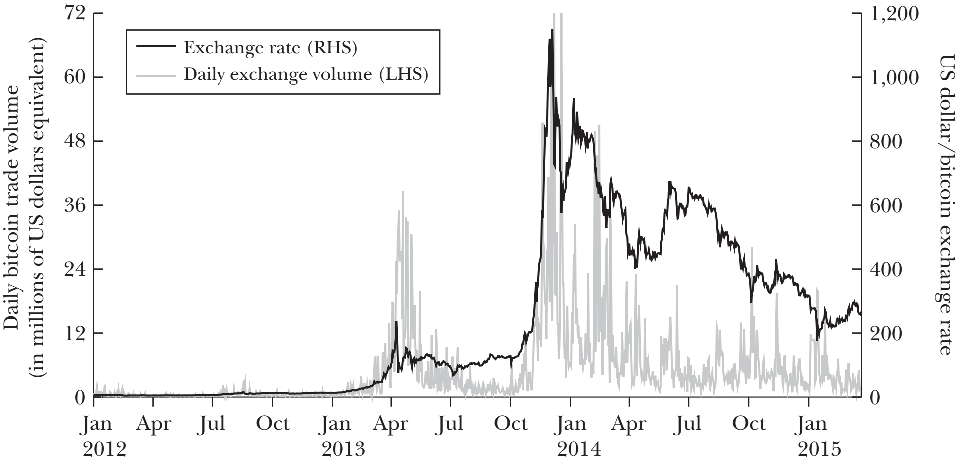
Bitcoin's design presents distinctive risks that differ from other payment methods and stores of value. Here, we review market risk, the shallow market problem, counterparty risk, transaction risk, operational risk, privacy-related risk, and legal and regulatory risks.

Any user holding bitcoins faces *market risk* via fluctuation in the exchange rate between bitcoin and other currencies. Figure 2 plots the average US dollar–bitcoin exchange rate at the largest exchanges, along with the weekly trade volumes. A user might dismiss the short-term price spikes before mid-2013 as part of the price of using a new currency. But the sharp movements from late 2013 through 2015 would be a source of concern, both for users considering Bitcoin for transactions and for those using it as a store of value.

The relatively low weekly trade volumes suggest that Bitcoin users also experience a *shallow markets problem*: for example, a person seeking to trade a large amount of bitcoin typically cannot do so quickly without affecting the market price.

Given centralization in the Bitcoin ecosystem, *counterparty risk* has become substantial. Exchanges often act as de facto banks, as users convert currency to bitcoin but then leave the bitcoin in the exchange. However, 45 percent of the Bitcoin currency exchanges studied by Moore and Christin (2013) ultimately ceased operation. High-volume exchanges were more likely to close because of a security breach, while operators of low-volume exchanges were more likely to abscond without explanation. Of the exchanges that closed, 46 percent did not reimburse their customers after shutting down. If users avoid holding their bitcoins in an exchange and instead use a digital wallet service, other risks arise, as these firms

Figure 2
US Dollar–Bitcoin Exchange Rate, January 2012–March 2015, along with Daily Bitcoin Trade Volume (in US Dollar Equivalent) at Four Top Currency Exchanges



Source: Authors using data from Blockchain.info and Quandl.com.

have become a lucrative target for cybercriminals. Examples include 4,100 bitcoins (valued at $1.2 million at then-applicable rates) taken from Bitcoin wallet inputs.io in November 2013, leading to that company's default (McMillan 2013) as well as 1,295 bitcoins ($1 million) taken from Bitcoin payment processor BIPS the next month following denial-of-service attacks (Southurst 2013).

The irreversibility of Bitcoin payments creates heightened *transaction risk*. If bitcoins are sent due to error or fraud, the Bitcoin system offers no built-in mechanism to undo the error. Of course, a buyer and seller can voluntarily agree to correct errors, but the Bitcoin protocol has no mechanism to retake the funds by force. In a world of competing payment methods, irreversibility puts Bitcoin at a disadvantage: all else equal, consumers should favor a payment system that allows reversal of unwanted or mistaken charges.

Transaction risk also arises when receiving payments. As discussed above, Bitcoin transactions do not clear (and hence are not final) until they have been added to the authoritative block chain. Transaction batches are only added every ten minutes on average. This creates at least two potential avenues for abuse. First, there is a low but persistent risk that what was once viewed as the authoritative block chain will later be cast aside, as voted on by a majority of participants, canceling any transactions recorded in that version of the block chain. Second, malevolent participants could double-spend bitcoins, particularly through rapid transactions before the block chain is updated. The protocol has taken steps to mitigate this possibility, but researchers have demonstrated viable attacks if Bitcoin is used for faster payments than intended by design (Karame, Androulaki, and Čapkun 2012).

A separate transaction risk arises from proposals to blacklist tainted Bitcoins, specifically those that have been obtained through theft. Some set of arbiters would publicly announce the ill-gotten bitcoins (much like a list of serial numbers on stolen paper currency), and the proposals call on the community to refuse incoming payments appearing on the blacklist. However, blacklists are controversial within the Bitcoin community (Bradbury 2013). After all, blacklists create the prospect of rejecting transactions that have already occurred—transferring losses to those who had unknowingly accepted bitcoin that later turned out to be ill-gotten. Blacklists add significant complexity and create a risk of abuse by those who manage the blacklists. Finally, widespread use of blacklists could undermine the fungibility of bitcoins. With the block chain available for public inspection, each bitcoin can be traced to its unique transaction history, and in principle market participants could place varying values on bitcoins according to their apparent risk of future blacklisting.

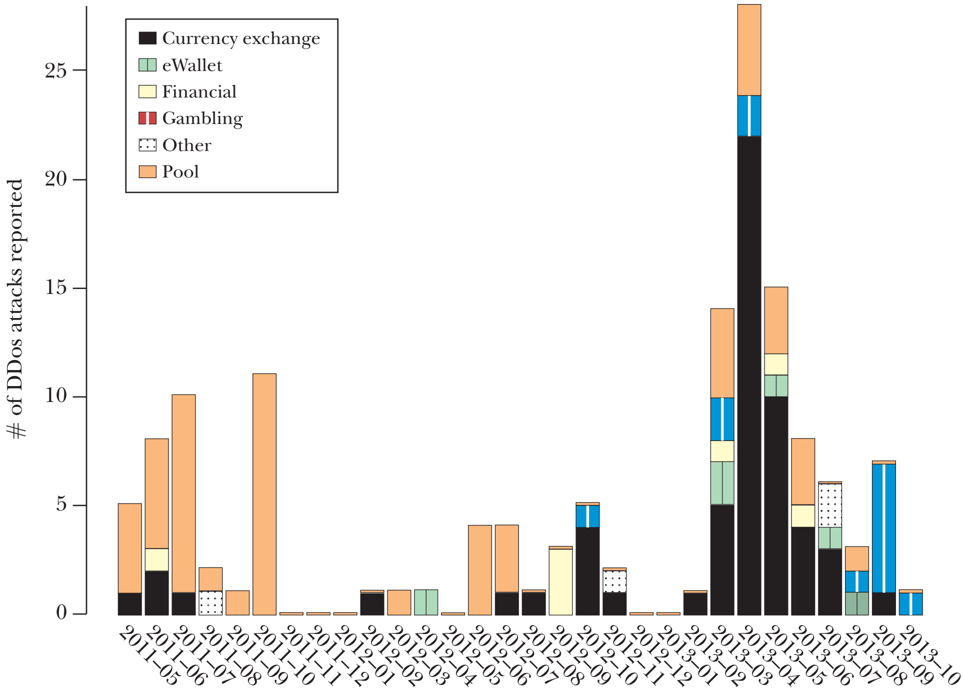
Operational risk encompasses any action that undermines Bitcoin's technical infrastructure and security assumptions. For example, despite a user's efforts to keep a private key secure, vulnerabilities are to be expected—including operator error, security flaws, and malware that scours hard drives in search of wallet credentials and private keys. At least as worrisome, the Bitcoin platform faces operational risks through potential vulnerabilities in the protocol design or breakthroughs in cryptanalysis. Community attention has focused on the so-called "51 percent attack," in which if some group can reliably control more than half the computational power, they can seize control of the system (Barber, Boyen, Shi, and Uzen 2012). If such attacks arose, the Bitcoin community might devise defenses, but the conflict and transition would be chaotic and would probably undermine trust in Bitcoin.

Denial-of-service attacks form a particularly prominent operational risk, particularly large for those who use Bitcoin through various intermediaries. Denial-of-service attacks entail swamping a target firm with messages and requests in such volume that it becomes unusable or very slow. Such attacks have diverse motivations. For example, an attack on a mining pool can prevent a pool's participants from solving the current puzzle and thus give an advantage to all other miners (Johnson, Laszka, Grossklags, Vasek, and Moore 2014). News of an attack can undermine trust in an exchange or even in Bitcoin itself—allowing an attacker to buy bitcoin at lower prices. Finally, attackers can demand ransom from service providers (such as exchanges), threatening attacks that would undermine the service's operation and customers' confidence. Figure 3 plots the number of denial-of-service attacks reported by users on the popular bitcointalk.org forum in 2011 to 2013, showing progression from attacks on mining pools to attacks on exchanges. While denial-of-service attacks occur throughout the web, they seem to be particularly effective in the Bitcoin ecosystem due to the relative ease of monetizing the attacks.

Bitcoin raises certain *privacy risks*, most notably the risk that transactions can be linked back to the people who made them. Bitcoin transactions are not truly anonymous: instead, they are *pseudonymous*, in that each transaction specifies account information (the user's public key) albeit without personal names, and the block

Figure 3

Reported (Distributed Denial of Service) DDoS Attacks on Bitcoin Services over Time



Source: Vasek, Thornton, and Moore (2014).

chain publishes transactions by that user identifier. Moreover, transactions made using Bitcoin often reveal real names—for example, as funds are converted to or from currencies in traditional banks, or when purchases from retailers reveal a customer name and mailing address. In principle, a Bitcoin user's identity could be obtained from one such source and then associated with the user's other transactions—flouting the widespread expectation of privacy.

Finally, Bitcoin systems face numerous *legal and regulatory risks* across countries. For example, a law-abiding user could lose funds in an exchange that is frozen or seized due to criminal activity—even if only a portion of the exchange's customers were in fact engaged in such activity. Furthermore, uncertain tax treatment of Bitcoin gains and losses hinders tax planning. We explore these questions in the next section.

Regulating Virtual Currencies

The original vision of Bitcoin is broadly in tension with regulation and government control. In this respect Bitcoin extends a line of cyber-libertarianism, traced back at least to John Perry Barlow's 1996 "Declaration of the Independence of Cyberspace," denying the role of governments in overseeing online communications. But contrary to the initial view that Bitcoin's decentralization made it impossible to regulate, there now appears to be ample possibility of regulatory oversight, as well as circumstances in which such intervention could be useful.

Fighting Crime

Bitcoin receives regulatory scrutiny for three classes of criminal concerns: Bitcoin-specific crime, money laundering, and Bitcoin-facilitated crime.

Bitcoin-specific crimes are attacks on the currency and its infrastructure like bitcoin theft, attacks on mining pools, and denial-of-service attacks on exchanges to manipulate exchange rates. Law enforcement often struggles to prevent or solve these crimes due to their novelty, lack of clarity on which agency and jurisdiction are responsible, technical complexity, procedural uncertainty, and limited resources.

Second, Bitcoin can be used for *money laundering*. Bitcoin money laundering could evolve to become more difficult to trace, particularly when funds are routed through mixers, with mixing records concealed from the public and perhaps unavailable to law enforcement. These characteristics might assist perpetrators in concealing or mischaracterizing the proceeds of crime. That said, Bitcoin also includes design elements that could facilitate the tracing of funds, including publication of the block chain (providing permanent publicly available records of what funds moved where).

Finally, *Bitcoin-facilitated crime* entails payment for unlawful services delivered (or purportedly delivered) offline, like the illegal goods and services sold on Silk Road and payment of funds in extortion. Criminals may be drawn to virtual currencies because they perceive a lack of regulatory oversight, because they distinctively value irreversible transactions, or because they have been banned or ejected from other payment mechanisms.

Consumer Protection

A related justification for regulatory action is the need for consumer protection. Such discussions were particularly frequent after the February 2014 failure of Bitcoin exchange Mt. Gox, which lost bitcoins valued at more than $300 million. In light of this failure and others (Moore and Christin 2013), it is desirable to have orderly processes that distribute any remaining assets equitably. The risk of collapse also calls for disclosures to help consumers understand the products they are buying.

Broader consumer protection concerns result from irreversibility of Bitcoin transfers. Most electronic payment systems provide mechanisms to protect consumers against unauthorized transfers, and indeed such protections are often codified into law. (For example, credit card dispute rights are guaranteed by the US Fair Credit and Billing Act, 15 USC § 1666.) The absence of such protections in Bitcoin therefore appears to be contrary to longstanding public policy.

Regulatory Options

A key challenge for prospective regulators is where to impose constraints. It is infeasible to regulate all peers in the Bitcoin network due to their quantity, their geographic distribution, and the privacy protections in the network. Instead, regulators are naturally drawn to key intermediaries. But intermediaries raise predictable defenses. Why, they ask, should they face liability for the conduct of third-party users, customers, or suppliers? Furthermore, some users will anticipate regulators targeting intermediaries and will act to avoid such scrutiny, just as criminals can pay each other in cash to hide illegal activities from financial institutions.

The FBI takedown of Silk Road in 2013 illustrates both the challenges of regulation and regulators' ultimate power. Silk Road was hosted as a "hidden service" on Tor, a system which is purpose-built for anonymity of both visitors and operators. Payments were only accepted in bitcoin. However, the Silk Road domain site was seized by the FBI when the site's alleged operator, Ross Ulbricht, was arrested on charges of conspiracy to distribute controlled substances, computer hacking, money laundering, and murder-for-hire charges. The private keys associated with Ulbricht's 144,000 bitcoins were also seized by the FBI (Greenberg 2013). Investigators targeted large merchants and administrators on Silk Road, exploiting poor operational security tactics to find their real identities. Ulbricht himself was identified by finding an early Silk Road advertisement posted on an online forum using his personal Gmail address (Zetter 2013). Silk Road's online presence and electronic records in some respects made it an easier target than, say, a small-time dealer of drugs or weapons.

Transfers through currency exchanges are also within regulators' grasp. In March 2013, the US Financial Crimes Enforcement Network issued guidance on when virtual currency operators should be classified as money-services businesses, requiring certain registration, reporting, and recordkeeping obligations. As exchanges complied, account details became available to regulators, and two months later, a US judge signed a seizure warrant for an account at the Mt. Gox exchange. In China, a December 2013 policy was broadly similar, requiring that Bitcoin intermediaries implement know-your-customer registrations for account-holders (People's Bank of China and Five Associated Ministries 2013). These regulatory requirements will not impede peer-to-peer bitcoin transactions that are not facilitated by currency exchanges. But it seems longstanding reporting requirements can provide a level of compliance for virtual currencies similar to what has been achieved for traditional currencies.

In principle, Bitcoin's electronic implementation in some ways makes it easier to regulate than offline equivalents. Consider the problem of theft. Once stolen cash enters circulation, little can be done to reclaim it. In contrast, Bitcoin blacklists could let law enforcement claw back all ill-gotten or stolen bitcoins—albeit with the problems discussed earlier.

Tax treatment of Bitcoin remains unsettled. In March 2014, the Internal Revenue Service (2014) issued guidance that transactions to and from virtual currencies may create taxable events for federal tax purposes. Thus, if a user converts dollars to bitcoin at one exchange rate, then later converts back at a higher rate, the user may owe tax on the appreciation; conversely, losses could offset gains elsewhere. Depending on the user's purpose and primary activity, the gains and losses could be ordinary income or capital (Notice 2014-21). While this guidance seems well-grounded in longstanding principles of US tax law, it was criticized for creating additional record-keeping and complexity, particularly for those whose conversions are frequent.

While Bitcoin now appears to be subject to regulatory oversight, the authority of regulators faces certain limits. For example, if one country places too large a burden on Bitcoin services based there, services are likely to develop elsewhere. If many countries impede use of Bitcoin, some users will resort to services like Zerocash with even stronger security precautions—likely letting criminals continue to use the service yet, perhaps, adding too much complexity for mainstream consumers. The overall regulatory goal should not take aim at Bitcoin or any other specific system or company, but instead should consider regulations in the broader context of a global market for virtual currency services.

Bitcoin as a Social Science Laboratory

Bitcoin has the potential to be a fertile area for social science research. Scholars should appreciate Bitcoin's contained environment with a clear set of rules (albeit not free from frictions), the publicly available record of transactions (unusual for most means of exchange), and the general availability of data even beyond the block chain (including market prices and trading volumes). To date, researchers have considered diverse questions ranging from design of financial markets to user behavior along with myriad questions of law and regulation. This research is of course quite recent, and much of it is still in working paper form. Many questions remain open, particularly to researchers who combine a deep understanding of Bitcoin with technical skills to collect data and a solid background in social science. Here are some of the issues this research has tackled and could approach in the future.

Bitcoin as a Financial Asset

After comparing exchange-traded volume of bitcoins to total transaction volume within the Bitcoin network, Glaser, Zimmerman, Haferkorn, Weber, and Sterling (2014) conclude that most users (by volume) treat their bitcoin investments as speculative assets rather than as means of payment. Bitcoin investments seem to offer diversification benefits according to Brière, Oosterlinck, and Szafarz (2013), who study correlations between bitcoin and other asset classes. Gandal and Halaburda (2014) examine exchange rates of different virtual currencies to observe comovement and identify opportunities for triangular arbitrage. Preliminary results on daily "closing" prices indicate little opportunity, although this may reflect that the arbitrageurs operate faster than the frequency of data points. Of course, given ongoing fluctuations in bitcoin prices and innovations in other virtual currencies, new data is already available for these kinds of studies.

Incentive-compatibility in Bitcoin Protocols

When confronted by a set of protocols, economic agents naturally look for ways to participate that increase their own gains. For example, early mining pools faced selfish behavior in the form of "pool hopping": Miners opted out of the pool in long rounds, in which the potential block reward has to be shared with a larger group. This drew attention to the mechanism design problem of keeping the expected payoff constant over time (Rosenfeld 2011).

Overall, the standard Bitcoin client software does not always act in the best interest of its principal. Both on the peer-to-peer network layer (Babaioff, Dobzinski, Oren, and Zohar 2012) and for the block mining protocol (Eyal and Sirer 2014), the prescribed rules are not equilibrium strategies if one considers the option to withhold information on a selective and temporary basis. Furthermore, Houy (2014b) observes that larger blocks are less likely to win a block race than smaller ones, meaning that a miner reduces the chance of collecting a reward when including new transactions into blocks—raising the question of why miners include transactions into blocks at all. So far, these concerns are theoretical. We are not aware of empirical evidence demonstrating substantial deviations from the suboptimal rules.

Privacy and Anonymity

The protection of online privacy and personal information arises in many contexts, and Bitcoin offers a specific set of rules and firms like the "mixers" that seek to offer privacy—although as we have seen, the privacy protections can be breached in various ways. Several papers analyze the public Bitcoin transaction history (Reid and Harrigan 2012; Ober, Katzenbeisser, and Hamacher 2013; Ron and Shamir 2013), finding a set of heuristics that can help to link Bitcoin accounts with real-world identities as long as some additional information is available for a related transaction. Androulaki, Karame, Roeschlin, Scherer, and Čapkun (2013) quantify the anonymity in a simulated environment similar to Bitcoin, finding that almost half of the users can be identified by their transaction patterns.

Monetary Policy

In a broad sense, the Bitcoin economy implements a variant of Milton Friedman's (1960, p. 90) "k-percent rule"—that is, a proposal to fix the annual growth rate of the money supply to a fixed rate of growth. Indeed, Bitcoin's protocol calls for an end of the minting phase at which point $k = 0$. In fact, k may even be negative in the future, because bitcoins can be irreversibly destroyed when users forget their private keys. This raises one of the classic questions in monetary policy: What happens when the size of an economy grows at a different rate than the quantity of money in that economy? Or if viewing Bitcoin as a social science laboratory, what happens if the Bitcoin economy grows faster than the supply of bitcoins?

Just as overly rapid growth of a money supply is classically linked to inflation, the fixed slow growth rate of Bitcoin creates the possibility of deflation if Bitcoin was to be used widely, as Krugman (2011) noted while comparing the Bitcoin economy to the gold standard. In response to this risk, developers proposed alternative system rules. For example, Primecoin and Peercoin modify Bitcoin to provide an unlimited money supply, with *k* fixed to approximately 1 percent for Peercoin.

It remains unclear whether decentralized cryptographic currencies can be designed with monetary policies that include feedback or even discretion. Bitcoin's design embodies a basic version of monetary policy that does not consider the state of the real economy. We note that Bitcoin's block chain presents a crude measure of monetary indicators—specifically the number of transactions and their nominal amount—but offers no information about what value was actually provided in exchange for payment. The block chain thus lays the groundwork for automatic monetary policy based solely in nominal data, but does not facilitate any policy based on real economic activity. Human arbiters could presumably add information about economic conditions or could introduce discretion by judgment, but they would also introduce the governance questions Bitcoin set out to overcome. Further experience with Bitcoin and other virtual currencies may illuminate some of the longstanding issues on the conduct and effects of monetary policy.

Looking Ahead

What is the future of Bitcoin and other virtual currencies? To replace credit cards for everyday consumer payments? To displace Western Union and other firms for international cash payments? To supplant banks for short-term deposits? Will Bitcoin and other virtual currencies favor low costs (to undercut competitors), privacy (to serve users who distinctively seek that benefit), or decentralization (to avoid a single point of control)? When disputes arise, do Bitcoin service providers protect sellers (who seek finality) or buyers (who often want refunds)? The original vision of Bitcoin offered one set of answers, but as new constituents approach the service, it becomes less clear that early design decisions meet prevailing requirements. It is also uncertain whether a single service can serve all needs. For example, those who seek greater privacy may be prepared to accept greater technical complexity and perhaps higher fees. However, recruiting mainstream consumers and merchants seems to call for a focus on simplicity and lower prices.

Bitcoin may be able to accommodate a community of experimentation built on its foundations. Mixers already close the most obvious privacy shortcomings in Bitcoin's early design, while pools help reduce risk for miners, and wallets address some of consumers' usability and security concerns.

Other aspects of Bitcoin architecture are largely locked in place through its protocol design. For example, the block chain is the essence of Bitcoin. There is no clear way for Bitcoin to substitute a different approach to record-keeping while retaining installed Bitcoin software, remaining compatible with intermediary systems, and, most importantly, retaining the overall consensus that has coordinated around Bitcoin. Instantaneous transaction confirmations seem to require equally fundamental changes. In these and other respects, Bitcoin will struggle to make adjustments.

Numerous competing virtual currencies are waiting in the wings. For example, Litecoin confirms transactions four times faster than Bitcoin, potentially facilitating retail use and other time-sensitive transactions. NXT reduces the electrical and computational burden of Bitcoin mining by replacing proof-of-work mining with proof-of-stake, assigning block chain duties in proportion to coin holdings. Zerocash (Ben-Sasson et al. 2014), which is not yet operational, will seek to improve privacy protections by concealing identifiers in the public transaction history. Peercoin allows a perpetual 1 percent annual increase in the money supply.

To offer their competing design decisions, alternative virtual currencies would first need to achieve confidence in their value and adoption. Bitcoin benefited from early excitement for its service, buyers and sellers at Silk Road, and favorable press coverage. A replacement virtual currency would struggle to obtain this combination of advantages, but without favorable expectations for growth, few would be willing to convert traditional currency into a competing coin. Whether or not Bitcoin expands as its proponents envision, it offers a remarkable experiment, a lab for researchers, and an attractive means of exchange for a subset of merchants and consumers.

References

Abrams, Rachel, Goldstein, Matthew, and Tabuchi, Hiroko. 2014. "Erosion of Faith Was Death Knell for Mt. Gox." *New York Times*, February 28.

Anderson, Ross. 2012. "Risk and Privacy Implications of Consumer Payment Innovation in the Connected Age." Proceedings of the Conference on Consumer Payment Innovation in the Connected Age, Federal Reserve Bank of Kansas City, March 29–30. pp. 99–116. http://www.kc.frb.org/publicat/pscp/2012/Session-3.pdf.

Andreessen, Marc. 2014. "Why Bitcoin Matters." *New York Times,* DealBook, January 21.

Androulaki, Elli, Ghassan O. Karame, Marc Roeschlin, Tobias Scherer, and Srdjan Čapkun. 2013. "Evaluating User Privacy in Bitcoin." In *Financial Cryptography and Data Security*, vol. 7859 of *Lecture Notes in Computer Science*, pp. 34–51. Springer.

Andrychowicz, Marcin, Stefan Dziembowski, Daniel Malinowski, and Łukasz Mazurek. 2014. "Secure Multiparty Computations on Bitcoin." *Proceedings of the 35th IEEE Symposium on Security and Privacy*, May 18–21. IEEE Press.

Babaioff, Moshe, Shahar Dobzinski, Sigal Oren, and Aviv Zohar. 2012. "On Bitcoin and Red Balloons." In *Proceedings of 13th ACM Conference on Electronic Commerce*, pp. 56–73. ACM.

Barber, Simon, Xavier Boyen, Elaine Shi, and Ersin Uzun. 2012. "Bitter to Better—How to Make Bitcoin a Better Currency." In *Financial Cryptography and Data Security*, vol. 7397 of *Lecture Notes in Computer Science*, pp. 399–414. Springer.

Barlow, John Perry. 1996. "A Declaration of the Independence of Cyberspace." https://projects.eff.org/~barlow/Declaration-Final.html (last accessed March 15, 2015).

Ben-Sasson, Eli, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. 2014. "Zerocash: Decentralized Anonymous Payments from Bitcoin." *Proceedings of the 2014 IEEE Symposium on Security and Privacy,* May 18–21, 2014.

Bitcoincharts.com. 2015. "Bitcoin Network." http://bitcoincharts.com/bitcoin/ (last accessed March 20, 2015).

Bitcoin Wiki. 2015a. "Category:Gambling." https://en.bitcoin.it/wiki/Category:Gambling (last accessed March 20, 2015).

Bitcoin Wiki. 2015b. "Mining Hardware Comparison." https://en.bitcoin.it/wiki/Mining_hardware_comparison (last accessed March 20, 2015).

Bitcoinity.org. 2015. "Exchanges." https://bitcoinity.org/markets/list (last accessed March 20, 2015).

Blockchain.info. 2015. "Bitcoin Charts." https://blockchain.info/charts/ (last accessed March 20, 2015).

Böhme, Rainer. 2013. "Internet Protocol Adoption: Learning from Bitcoin." *Proceedings of the IAB Workshop on Internet Technology Adoption and Transition (ITAT)*, Cambridge, UK.

Bonneau, Joseph. 2014. "Estimating the Power Consumption of Bitcoin." Presented at Financial Cryptography and Data Security, 18th International Conference (rump session), Bridgetown, Barbados, March 4, 2014.

Bradbury, Danny. 2013. "Anti-Theft Bitcoin Tracking Proposals Divide Bitcoin Community." *Coindesk*, November 15.

Brière, Marie, Kim Oosterlinck, and Ariane Szafarz. 2013. "Virtual Currency, Tangible Return: Portfolio Diversification with Bitcoins." Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2324780.

Chaum, David. 1983. "Blind Signatures for Untraceable Payments." In *Advances in Cryptology,* edited by D. Chaum, R. L. Rivest, and A. T. Sherman, 199–203. Springer.

Christin, Nicolas. 2013. "Traveling the Silk Road: A Measurement Analysis of a Large Online Anonymous Marketplace." *Proceedings of the 22nd International World Wide Web Conference (WWW'13),* pp. 213–24. Rio de Janeiro, Brazil, May 2013.

Clark, Jeremy, Joseph Bonneau, Edward W. Felten, Joshua A. Kroll, Andrew Miller, and Arvind Narayanan. 2014. "On Decentralizing Prediction Markets and Order Books." Workshop on the Economics of Information Security, State College, Pennsylvania, June 2014.

Clark, Jeremy, and Aleksander Essex. 2012. "CommitCoin: Carbon Dating Commitments with Bitcoin." In *Financial Cryptography and Data Security,* vol. 3797 of *Lecture Notes in Computer Science,* pp. 390–98. Springer.

Danezis, George, and Claudia Diaz. 2008. "A Survey of Anonymous Communication Channels." Microsoft Research Technical Report MSR-TR-2008-35.

Diffie, Whitfield, and Martin E. Hellman. 1976. "New Directions in Cryptography." *IEEE Transactions on Information Theory* 22(11): 644–54.

Dingledine, Roger, Nick Mathewson, and Paul Syverson. 2004. "Tor: The Second-Generation Onion Router." In *Proceedings of the 2004 USENIX Security Symposium.* USENIX.

D.K. 2013. "Bitcoin's Collapse: China Blues." *The Economist,* December 18.

Edelman, Benjamin. 2014. "Consumers Pay More When They Pay with Bitcoin." PYMNTS.com, May 20.

European Central Bank. 2012. *Virtual Currency Schemes.* Technical Report, October. https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemes201210en.pdf (last accessed April 1, 2015).

Evans, David S. 2014. "Economic Aspects of Bitcoin and Other Decentralized Public-Ledger Currency Platforms." Coase-Sandor Institute for Law and Economics Working Paper 685, April.

Eyal, Ittay, and Emin Gün Sirer. 2014. "Majority is Not Enough: Bitcoin Mining is Vulnerable." In *Financial Cryptography and Data Security*, vol. 8437 of *Lecture Notes in Computer Science,* pp. 436–54. Springer.

Financial Crimes Enforcement Network (FinCEN), US Department of the Treasury. 2015. "Enforcement Actions for Failure to Register as a Money Services Business." http://www.fincen.gov/news_room/ea/ea.msb.html (last accessed March 20, 2015).

Friedman, Milton. 1960. *A Program for Monetary Stability.* New York: Fordham University Press.

Gandal, Neil, and Hanna Halaburda. 2014. "Competition in the Crypto-Currency Market." Presentation at the Workshop on the Economics of Information Security, State College, PA, June 2014.

Glaser, Florian, Kai Zimmermann, Martin Haferkorn, Moritz Christian Weber, and Michael Siering. 2014. "Bitcoin—Asset or Currency? Revealing Users' Hidden Intentions." *Proceedings of the 22nd European Conference on Information Systems,* Tel Aviv, June 2014.

Greenberg, Andy. 2013. "FBI Says It's Seized $28.5 Million in Bitcoins from Ross Ulbricht, Alleged Owner of Silk Road." *Forbes,* October 25.

Hicks, John R. 1967. *Critical Essays in Monetary Theory.* Clarendon Press.

Houy, Nicolas. 2014a. "The Economics of Bitcoin Transaction Fees." GATE WP 1407

Université de Lyon, Groupe d'Analyse et de Théorie Economique (GATE), February. Available at SSRN: http://ssrn.com/abstract=2400519.

Houy, Nicolas. 2014b. "The Bitcoin Mining Game." March. Available at SSRN: http://ssrn.com/abstract=2407834.

Internal Revenue Service (IRS). 2014. "IRS Virtual Currency Guidance: Virtual Currency Is Treated as Property for U.S. Federal Tax Purposes; General Rules for Property Transactions Apply." March 25.

Johnson, Benjamin, Aron Laszka, Jens Grossklags, Marie Vasek, and Tyler Moore. 2014. "Game-Theoretic Analysis of DDoS Attacks against Bitcoin Mining Pools." In *Financial Cryptography and Data Security*, vol. 8438 of *Lecture Notes in Computer Science*, Part I: First Workshop on Bitcoin Research, pp. 72–86. Springer.

Karame, Ghassan O., Elli Androulaki, and Srdjan Čapkun. 2012. "Double-spending Fast Payments in Bitcoin." In *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS '12)*, pp. 906–917. ACM.

Kim, Victoria. 2014. "Dutch National Pleads Guilty to Running Online Marketplace for Drugs." *Los Angeles Times*, September 3.

Krugman, Paul. 2011. "Golden Cybervetters." *New York Times,* September 7.

MacCarthy, Mark. 2010. "What Payment Intermediaries Are Doing about Online Liability and Why It Matters." *Berkeley Technology Law Journal* 25(2): 1037–1120.

Matonis, Jon. 2013. "Bitcoin Casinos Release 2012 Earnings." *Forbes,* January 22.

McLeod, Andrew Saks. 2013. "Bitcoins Soar in Value in Argentina due to Capital Control Laws." *Forex Magnates,* July 9.

McMillan, Robert. 2013. "$1.2M Hack Shows Why You Should Never Store Bitcoins on the Internet." *Wired.* November 7.

Meiklejohn, Sarah, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. 2013. "A Fistful of Bitcoins: Characterizing Payments among Men with No Names." In *Proceedings of the 2013 ACM Internet Measurement Conference (IMC),* pp. 127–40. ACM.

Moore, Tyler, and Nicolas Christin. 2013. "Beware the Middleman: Empirical Analysis of Bitcoin-Exchange Risk." In *Financial Cryptography and Data Security*, vol. 7859 of *Lecture Notes in Computer Science*, pp. 25–33. Springer.

Möser, Malte, and Rainer Böhme. 2014. "Trends, Tips, Tolls: A Longitudinal Study of Bitcoin Transaction Fees." Available at SSRN: http://ssrn.com/abstract=2530843. Forthcoming in *Proceedings of the 2nd Workshop on Bitcoin Research*, January 2015.

Möser, Malte, Rainer Böhme, and Dominic Breuker. 2013. "An Inquiry into Money Laundering Tools in the Bitcoin Ecosystem." *Proceedings of APWG eCrime Researchers Summit*, San Francisco.

Nakamoto, Satoshi. 2008. "Bitcoin: A Peer-to-Peer Electronic Cash System." https://bitcoin.org/bitcoin.pdf (last accessed March 31, 2015).

Ober, Micha, Stefan Katzenbeisser, and Kay Hamacher. 2013. "Structure and Anonymity of the Bitcoin Transaction Graph." *Future Internet* 5(2): 237–50.

People's Bank of China and Five Associated Ministries. 2013. "Prevention of Risks Associated with Bitcoin." Notice. Translation available at https://vip.btcchina.com/page/bocnotice2013 (last accessed March 20, 2015).

Raskin, Max. 2013. "Meet the Bitcoin Millionaires." *Bloomberg Businessweek,* April 10.

Reid, Fergal, and Martin Harrigan. 2012. "An Analysis of Anonymity in the Bitcoin System." In *Security and Privacy in Social Networks,* edited by Yaniv Altshuler et al., pp. 197–223. Springer.

Ron, Dorit, and Adi Shamir. 2013."Quantitative Analysis of the Full Bitcoin Transaction Graph." In *Financial Cryptography and Data Security*, vol. 7859 of *Lecture Notes in Computer Science*, pp. 6–24. Springer.

Rosenfeld, Meni. 2011. "Analysis of Bitcoin Pooled Mining Reward Systems." November 17. https://bitcoil.co.il/pool_analysis.pdf (last accessed March 20, 2015).

Rosenfeld, Meni. 2012. "Overview of Colored Coins." https://bitcoil.co.il/BitcoinX.pdf (last accessed March 20, 2015).

Sidel, Robin. 2014. "Overstock CEO Sees Bitcoin Sales Rising More than Expected." *Wall Street Journal,* March 4.

Song, Sophie. 2014. "The Rise and Fall of Bitcoin in China: Central Bank Shuts Down All Chinese Bitcoin Exchanges." *International Business Times*, March 27.

Southurst, Jon. 2013. "Bitcoin Payment Processor BIPS Attacked, Over $1m Stolen." *CoinDesk,* November 25.

Taylor, Michael Bedford. 2013. "Bitcoin and the Age of Bespoke Silicon." Presented at the "International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES)," September–October, 2013. http://cseweb.ucsd.edu/~mbtaylor/papers/bitcoin_taylor_cases_2013.pdf (last accessed March 20, 2015).

United States of America v. Mark Peter Williams et al. 2011. C.D.CA Case No. CR-11-01137.

United States of America v. Ross William Ulbricht. Indictment. S.D.NY Case No. 14-CRIM-068.

United States of America v. Ross William

Ulbricht. Government Exhibit 940. S.D.NY Case No. 14-CRIM-068.

Vasek, Marie, Micah Thornton, and Tyler Moore. 2014. "Empirical Analysis of Denial-of-Service Attacks in the Bitcoin Ecosystem." In *Financial Cryptography and Data Security*, vol. 8438 of *Lecture Notes in Computer Science*, Part 1: First Workshop on Bitcoin Research, pp. 57–71. Springer.

World Nuclear Association. 2015. "Nuclear Power in the World Today." http://www.world-nuclear.org/info/Current-and-Future-Generation/Nuclear-Power-in-the-World-Today/ (last accessed March 20, 2015).

Yeow, Andy. 2015. "Global Bitcoin Nodes Distribution." https://getaddr.bitnodes.io/ (last accessed March 20, 2015).

Zetter, Kim. 2013. "How the Feds Took Down the Silk Road Drug Wonderland." *Wired*, November 18.

Journal of Economic Perspectives—Volume 29, Number 2—Spring 2015—Pages 239–258

Systematic Bias and Nontransparency in US Social Security Administration Forecasts[†]

Konstantin Kashin, Gary King, and Samir Soneji

Since the passage of the Social Security Act of 1935, a central concern of the Board of Trustees has been the demographic and financial forecasts necessary to assess the long-term solvency of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds. These forecasts are used in a variety of ways. For example, they affect decisions about whether the rate of payroll taxes or the amount of benefits should be raised or lowered for 212 million workers and 59 million beneficiaries in 2014, respectively. The methodology rooted in the forecasts is used by the Social Security Administration to evaluate policy proposals put forward by Congress to modify the program. The forecasts are also used as essential inputs in assessing the finances of Medicare and Medicaid and are central to research in demography, economics, political science, public health, public policy, and sociology. Although the Social Security Administration has performed these forecasts since 1942, no systematic and comprehensive evaluation of their accuracy has ever been published.

Each year, the Office of the Chief Actuary of the Social Security Administration carries out the mandate for producing forecasts in the *Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance*

■ *Konstantin Kashin is a PhD Candidate, Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts. Gary King is Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts. Samir Soneji is Assistant Professor, Dartmouth Institute for Health Policy & Clinical Practice, Dartmouth College, Lebanon, New Hampshire. Their email addresses are kkashin@fas.harvard.edu, king@harvard.edu, and samir.soneji@dartmouth.edu.*

[†] To access the Appendix and Data Appendix, visit http://dx.doi.org/10.1257/jep.29.2.239

doi=10.1257/jep.29.2.239

Trust Funds, commonly known as the "Trustees Report." Actuaries at the Social Security Administration (SSA) separately forecast demographic variables (for example, mortality rates) and economic variables (for example, labor force participation rates) that ultimately combine to produce solvency forecasts. In this article, we offer the first evaluation of Social Security forecasts that compares the SSA forecasts with observed truth; for example, we look at forecasts made in the 1980s, 1990s, and 2000s with outcomes that are now available. We do this first for demographic forecasts then for financial forecasts.[1]

Forecasts, of course, should not be expected to be precisely accurate. However, our analysis reveals several problems. First, Social Security Administration forecasting errors—as evaluated by how accurate the forecasts turned out to be—were approximately unbiased until 2000 and then became systematically biased afterward, and increasingly so over time. Second, most of the forecasting errors since 2000 are in the same direction, consistently misleading users of the forecasts to conclude that the Social Security Trust Funds are in better financial shape than turns out to be the case. Finally, the Social Security Administration's informal uncertainty intervals appear to have become increasingly inaccurate since 2000. Although the Social Security Administration has recently begun to follow the recommendations of its panel of outside technical advisers on including certain types of more formal uncertainty estimates, a step that should be part of all government reporting (Manski 2013), these estimates have also not been systematically evaluated.

At present, the Office of the Chief Actuary, at the Social Security Administration, does not reveal in full how its forecasts are made and, as a result, no other person, party, or organization, in or out of government, has been able to make fully independent quantitative evaluations of policy proposals about Social Security. Even the Congressional Budget Office, which produces Social Security Trust Fund solvency forecasts, relies on the demographic forecasts produced by the Office of the Chief Actuary as inputs for its models. Thus, the Office of the Chief Actuary holds an unusual position within American politics of being the sole supplier of Social Security forecasts, as well as heading the only organization producing fully independent quantitative evaluations of policy proposals to alter Social Security. For each evaluation of a proposed policy, the Office of the Chief Actuary estimates the effect on key financial outcomes that assess the solvency of the Trust Funds. For the vast majority of policy proposals evaluated by the Office of the Chief Actuary, the estimated financial impact is smaller than almost all of SSA's forecasting errors since 2000. Social Security Administration forecasts of current law and its counterfactual evaluation of policy proposals share the same growing bias because both are based on the same forecasting methodology. Additionally, the

[1] A replication dataset available with this article at http://e-jep.org and in Dataverse (Kashin, King, and Soneji 2015b) summarizes our data sources and all the information necessary to reproduce our results. An online Appendix, available at the same website, offers extra robustness checks.

uncertainty surrounding the estimated effects of proposed policies, which would likely be larger than the uncertainty in the forecasts under current law, usually dominate the estimated effect of the policy.

In the conclusion of the article, we argue that the Social Security Administration and its Office of the Chief Actuary should follow best practices in academia and many other parts of government and make their forecasting procedures public and replicable, and should calculate and report calibrated uncertainty intervals for all forecasts. In a companion paper, we offer an explanation for the origin of the biases reported here and propose simple structural ways of changing the system to fix the problems going forward (Kashin, King, and Soneji 2015a).

Demographic Forecasts

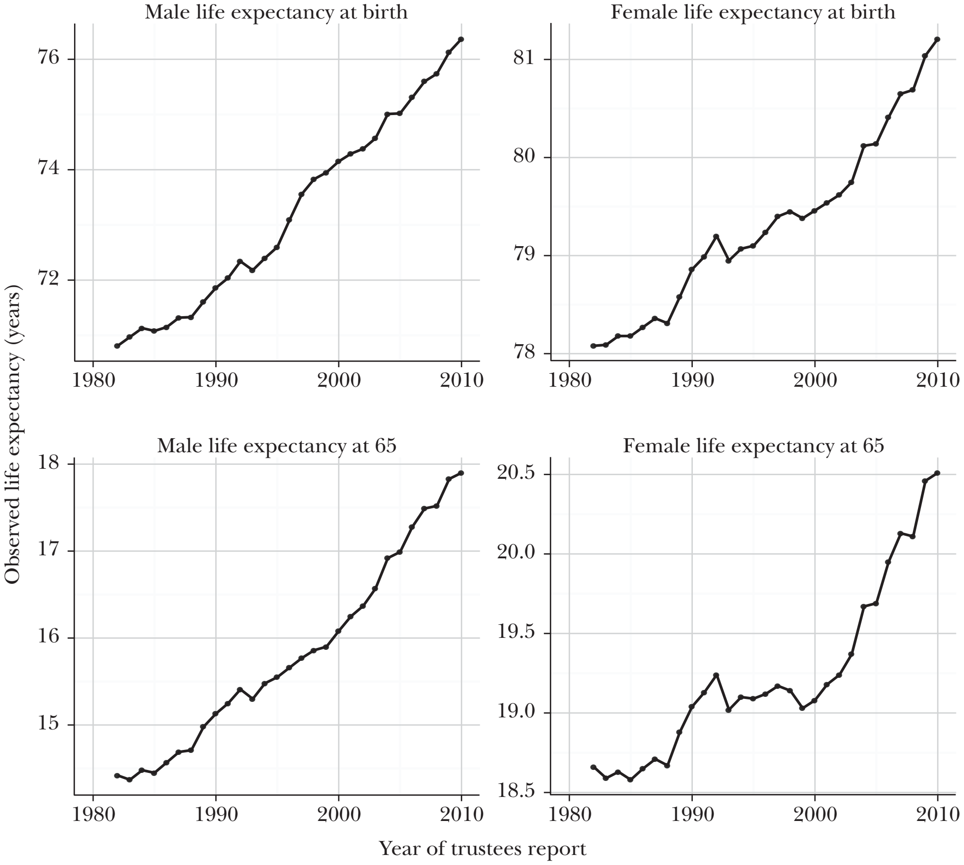
Demographic variables important to solvency forecasts from the Social Security Administration include mortality, fertility, and migration. Higher levels of fertility and migration increase the number of workers who contribute payroll taxes and increase long-term solvency. Lower levels of mortality, especially among those age 65 years and older, increase the number of retirees who receive benefits and decrease long-term solvency. Moreover, if Americans live longer than the forecasts predict, they will draw benefits for more years than expected and the Trust Funds will become exhausted sooner than anticipated. As Diamond and Orszag (2005, p. 63) explain, the increase in benefit payments from longer lives is not counterbalanced by an increase in payroll tax receipts because the system is designed to be approximately fair on average from an actuarial standpoint—the longer lives were not taken into account in what was paid into the system earlier in working life. People with longer working careers also receive higher benefits compared to those starting their careers at later ages.

Observed Demographic Data

As a baseline, we present four observed time series in Figure 1. Life expectancy for males and females, both at birth and at age 65, are relatively smooth over almost the entire time period, as can be seen in all four graphs. Indeed, three of the four are approximately linear; the fourth, female life expectancy at 65, is not far from linear. The highly regular nature of these data suggests that relatively accurate forecasts should be possible.

There is a standard conceptual difficulty in measuring current life expectancy: How can the analyst describe life expectancy when people are still alive? We follow common practice here by using the concept of "period life expectancy." This approach calculates life expectancy in a given year as the average number of years a person would expect to live if that person experienced the mortality rates in that given year over the course of a lifetime. Thus, life expectancy is a function of age-specific mortality rates and the average number of person-years contributed by those who die

Figure 1
Observed Period Life Expectancy



Note: "Period life expectancy" for a year is a single-number summary of all the age-specific mortality rates for that same year and is interpreted as the average number of years a person could expect to live if he or she experienced the mortality rates of a given year over the course of their life.

at each age. The mortality rate for people of a given age equals the number of deaths in that age divided by the number of person-years lived in that age (the exposure).[2]

The Office of the Chief Actuary forecasts male and female life expectancy separately. The male and female population counts are then combined with sex-specific economic factors like estimates of the labor force participation rates, and sex-specific beneficiary rates like disability incidence rates to project the population of workers and beneficiaries. In turn, the number of male and female workers

[2] We use observed life expectancy based on Human Mortality Database life tables rather than the life tables from the Social Security Administration. The small differences in estimated life expectancy between the two sources do not account for the much larger error rates and patterns reported in this article. Both sources seek to estimate the conditional probability of death (and life expectancy, its single

and beneficiaries serve as inputs for predictions of the operations and actuarial status of the Trust Funds.

We begin in 1982, the earliest year with regular life expectancy forecasts from the Social Security Administration, and continue until 2010, the last year for which observed actual data have been released. For the years before 2001, the Social Security Administration only reveals information about its demographic forecast for years divisible by five. However, the observed time series are quite smooth and, hence, interpolations to other years should be accurate.

Forecasts

We present our results in stages beginning in Figure 2 with an evaluation of forecasts for 2005 and 2010, the two years forecast by the largest number of Trustees Reports (with details for all years in our online Appendix). We compute forecast error (here and throughout) as the "intermediate scenario" forecast minus the observed value, so that positive values represent overestimates and negative values represent underestimates. The vertical axis is the forecast error for each of the four demographic variables, and the horizontal axis is the year of the Trustees Report when the forecast was made. Figure 2 illustrates four points.

First, despite the strong resemblance and very high correlation between male and female life expectancy in Figure 1, the forecast errors are substantially worse for males than females over most of the range of the forecasts. In some of the forecasts of the mid-1980s, the overestimate of female life expectancy is more-or-less offset by the underestimate of the male life expectancy, but in later years, both are underestimated.

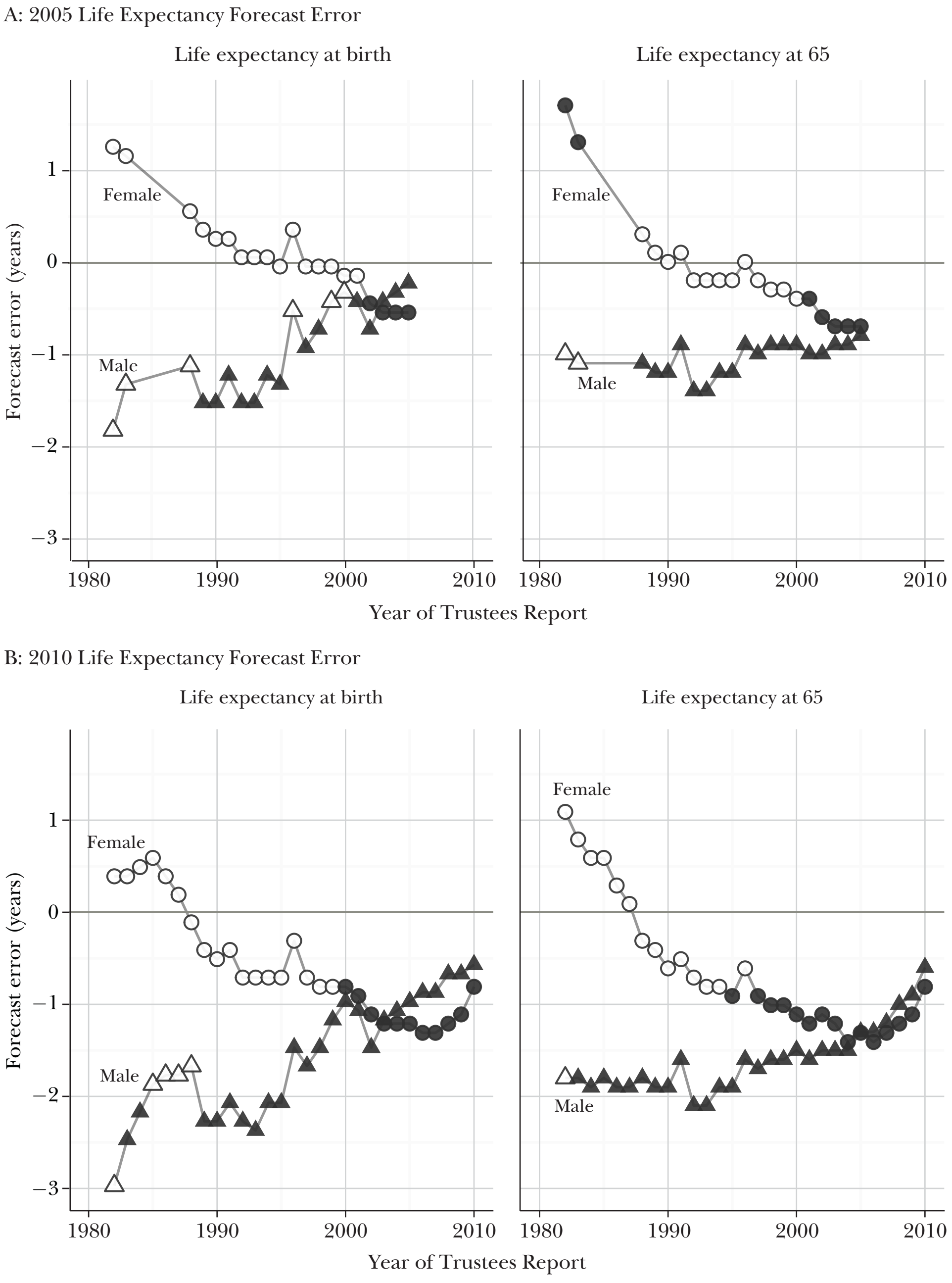
Second, the patterns of error persist. For example, Figure 2 shows that every single Trustees Report for 23 years from 1982–2005 underestimated male life expectancy in 2005. Similarly, every forecast for 28 years from 1982–2010 underestimated male life expectancy in 2010.

Third, the forecasts for males do pass an obvious test by more closely approximating the truth as the year being forecast approaches. For females, errors have been smaller than for males until recently, but in the years from 2000–2005, when projecting female life expectancy at 65, forecast errors of female life expectancy actually increased as the year of the Trustees Report approached the year being forecast.

Fourth, a large number of the forecasts fall outside the uncertainty intervals offered by the Social Security Administration. In the forecasts, these uncertainty intervals are categorized as "high cost," "intermediate cost," and "low cost" scenarios. The high- and low-cost scenarios form the Social Security Administration uncertainty interval. In Figure 2, we color points white if the truth falls within

number summary), but the Human Mortality Database is the standard in the scientific literature for its emphasis on "comparability, flexibility, accessibility, [and] reproducibility," for subjecting US Census counts and National Center for Health Statistics death counts to international quality standards, and for including all potential beneficiaries in the analyses, as discussed in an overview to the database (http://www.mortality.org/Public/Overview.php; accessed March 20, 2015). An online Appendix repeats all demographic analyses in this section with Social Security Administration data; no important differences arise compared with the results presented here.

Figure 2
Forecast Error of Life Expectancy in 2005 and 2010 by Year of Trustees Report



A: 2005 Life Expectancy Forecast Error

B: 2010 Life Expectancy Forecast Error

Note: The circles (females) and triangles (males) are colored white when truth falls within Social Security Administration uncertainty intervals and colored black when the truth falls outside these uncertainty intervals.

these "uncertainty intervals" and black if the truth falls outside of these intervals. Although any forecast is of course uncertain and errors are to be expected, uncertainty intervals should still capture the truth with some known frequency. We find that only one of the 29 uncertainty intervals for male life expectancy at age 65 for 2010 actually captured the true outcome. In the years after 2000, every single forecast for year 2010 male and female life expectancy at birth and at age 65 was underestimated, and the true outcome fell outside the uncertainty intervals.

The uncertainty intervals reported by the Social Security Administration are given no formal statistical basis in published materials, and thus we tried to assess how these intervals were qualitatively presented. In Trustees Reports from the earlier part of our period, the early and mid-1980s, the Social Security Administration wrote about the intervals as one would discuss wide confidence intervals, perhaps at a 90 percent confidence interval, and readers were warned that the confidence intervals might not necessarily cover the truth. In recent years, especially after 2000, the Trustees Reports became more confident in these intervals. Since 2003, the Trustees Reports have included an appendix referring to a stochastic model that attempts to formalize the uncertainty of their forecasts. The model itself is not publicly available, so outside analysts cannot evaluate how it has been calibrated or evaluated. But the Trustees describe the uncertainty intervals in qualitative terms that one would typically use to discuss something stronger than a 95 percent confidence level. For example, in the 2011, 2012, and 2013 Trustees Reports, the report repeated the same definition: "In the future, the costs of OASI, DI, and the combined OASDI programs as a percentage of taxable payroll are unlikely to fall outside the range encompassed by alternatives I [low cost] and III [high cost] because alternatives I and III define a wide range of demographic and economic conditions." (OASI refers to the old age and survivors program, which is commonly known as Social Security, while DI refers to the disability insurance program.)

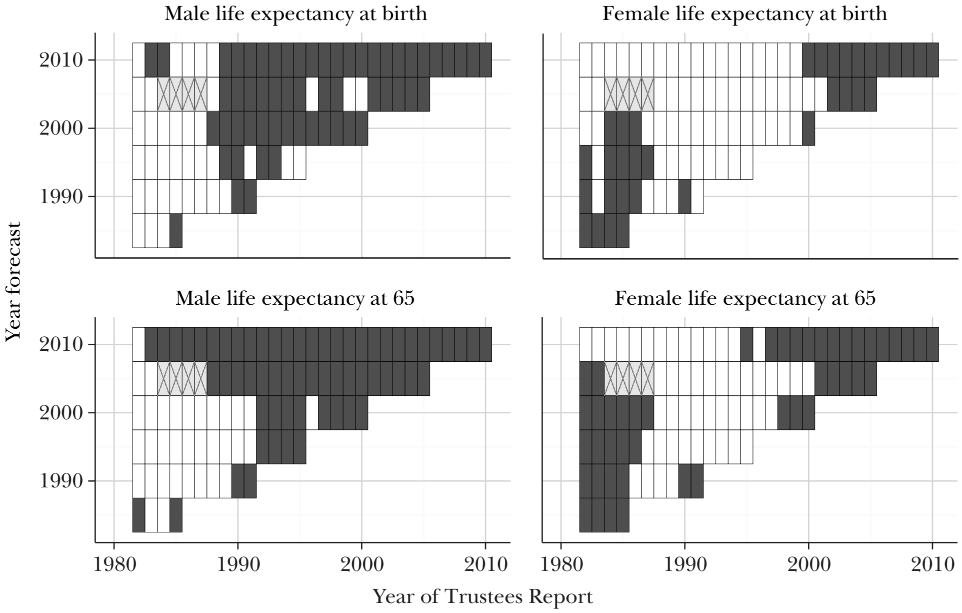
In short, the post-2000 forecasts all indicated that both men and women would have lived shorter lives than they did and also offered uncertainty ranges implying that the Trust Funds were on firmer financial ground than turned out to be warranted. We reach an identical conclusion when we examine the forecast error over all Trustees Reports for all observed years, as shown in the online Appendix.

Uncertainty Intervals

Finally, we analyze the set of all the Social Security Administration life expectancy forecasts with respect to uncertainty interval coverage. Figure 3 plots the year of the Trustees Report (horizontally) by the year of the forecast (vertically), with one square for each forecast colored white when the truth fell within the uncertainty interval and black when the truth fell outside the interval.

The results in Figure 3 demonstrate systematic problems with the uncertainty intervals used by the Social Security Administration. The uncertainty intervals failed to capture the truth for almost every forecast made since 2000 for all four demographic variables. For the graphs on male life expectancy at birth and at age 65 (the two graphs on the left), the problem began approximately in 1990.

Figure 3
Uncertainty Interval Coverage by Year of Trustees Report and Year of Forecast



Notes: White indicates uncertainty interval covered the truth, black indicates that it did not, and gray "X" indicates that the Social Security Administration did not provide an uncertainty interval. Contemporaneous forecast error is possible because of the time lag (typically three to four years) in finalizing mortality data.

We might expect that some uncertainty intervals fail to capture the eventually observed truth, especially when the forecast was made many years earlier than the year forecast. But since about 2000, the uncertainty intervals consistently failed to capture the truth for male and female life expectancy at birth and age 65. Apparently, the Social Security Administration did not perform a correction if and when these systematic errors became known.

Forecast Biases Are Not Explained by the Great Recession

Could the systematic forecasting biases documented in this section be caused to some extent by the Great Recession, which lasted from December 2007 to June 2009? Historically, increases in unemployment have led to lower mortality primarily because of fewer accidental deaths (like road traffic fatalities) (Granados 2005; Stuckler, Meissner, Fishback, Basu, and McKee 2011) not counterbalanced by a small increase in the comparatively fewer number of suicides. Thus, a lengthy recession could potentially explain life expectancies longer than predicted. But although the recession may explain some of the forecasting error, it cannot explain most of it.

First, the Great Recession began in December 2007, when the Social Security Administration had already been underestimating Americans' life expectancy for several years prior. Second, the mortality data and errors in forecasting mortality from one year to the next are relatively smooth functions of time—that is, the errors do not increase when the recession arrived. Finally, the Great Recession cannot account for the 0.6-year forecast error in male life expectancy and 0.8-year forecast error in female life expectancy made by the Social Security Administration in 2010. During the 18-month recession, unemployment increased 4.6 percentage points from a trough of 4.9 percent in February 2008 to a peak of 9.5 percent in June 2009. Previous US- and European-based studies estimate that mortality rates decline approximately 0.5 percent for every 1 percent in unemployment (Ruhm 2000). Thus, the 4.6 percentage point increase in employment during the Great Recession would correspond approximately to a 2.3 percent decline in mortality rates. For comparison, the inaccuracies in projected male and female life expectancies correspond to a 5.2 and 7.6 percent decline in mortality rates, respectively.

A Note about Fertility and Immigration

We also evaluated the performance of Social Security Administration forecasts of fertility and migration (with results shown in the online Appendix). Recent forecasts of the total fertility rate exhibited persistent and growing error, and the forecasts were overly confident. For example, the error in forecasts of the total fertility rate in 2010 grew—rather than shrank—across successive Trustees Reports. The forecast error of the 2010 total fertility rate in the 2010 Trustees Report was 0.15, which translated to approximately 315,000 more births forecasted than actually occurred (8 percent of total births in 2010).[3] As with mortality, forecast biases in fertility are not explained by the Great Recession. The rise in unemployment during the Great Recession led to a fertility decline of approximately 5 percent. Yet, the inaccuracy in the total fertility rate forecasted in 2010 corresponded to an approximately 8 percent difference in fertility rates. Overall, the forecast error in fertility makes the US population seemingly younger than it really is and, consequently, the Social Security Trust Funds healthier than they may be.

In contrast to mortality and fertility, Social Security Administration forecasts of legal immigration (the largest component of overall immigration) were far less biased and the confidence intervals seem appropriate. For example, the error in the forecast of net legal immigration in 2010 declined across successive Trustees Reports. By the 2010 Trustees Report, the forecast error was less than 1 percent of the observed number of net legal immigrants in 2010.

The results of mortality, fertility, and immigration forecasts may illuminate some of the reasons why the Social Security Administration varies in its performance of forecasting these three demographic components. As we discuss in detail in Kashin, King, and Soneji (2015a), a constellation of factors may have interacted to produce biased mortality and fertility forecasts. First, the forecasting method itself allows for the introduction of unintentional bias because it apparently involves a very large number (previously 210, now 150) of interrelated subjective decisions about rates

[3] The 2010 Trustees Report included historical fertility up to 2006 because of the time lag in reporting final birth data.

of mortality decline. Second, as Social Security reform has become more politically charged, the Social Security Administration seems to have disregarded the continued advice of its outside technical advisers to assume a more rapid increase in life expectancy. Third, mortality rates decreased at an ever faster pace after about 2000, but the Social Security Administration mortality forecasts did not keep pace with this change in input data.

Some of the same factors that possibly produce biased mortality forecasts may occur for fertility, too. The Social Security Administration forecasting method for fertility also involves subjective decisions about future levels of fertility rates. In contrast to mortality and fertility, the level of legal immigration is annually set by Congress. The Social Security Administration forecast of net legal immigration largely follows this Congressionally-set level; at present, it sets the ratio of legal emigration to legal immigration to 25 percent.

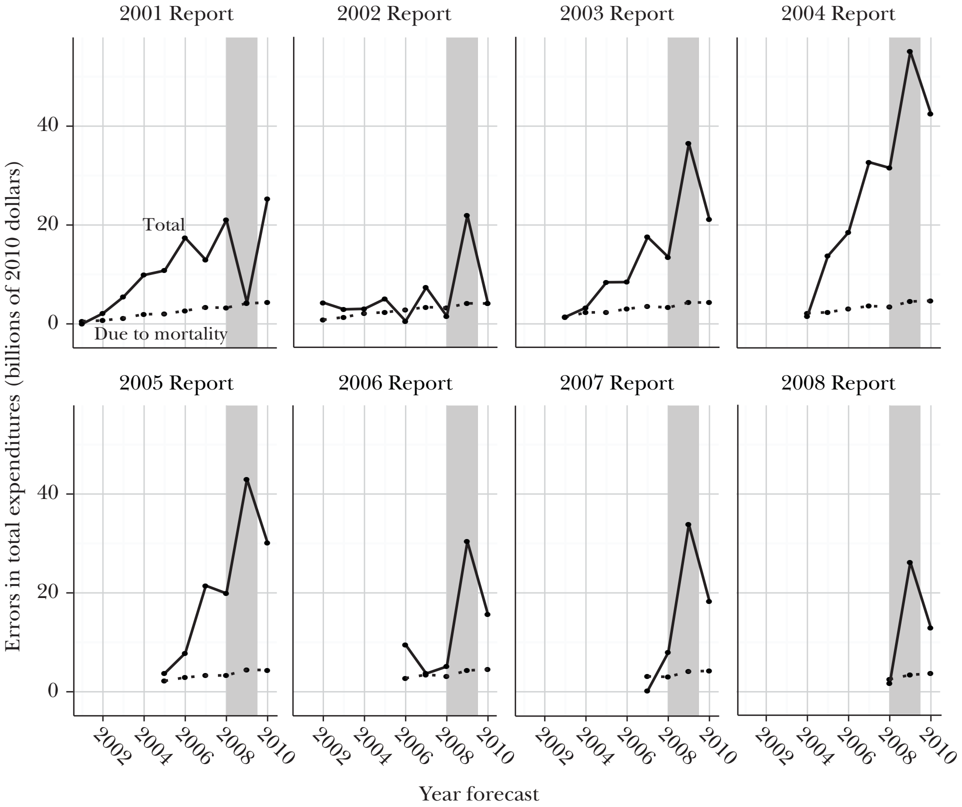
Financial Forecasts

We next consider Social Security Administration forecasts of Trust Fund solvency, for which demographic forecasts serve as a key input. In particular, we examine forecasts and observed outcomes of the three most commonly cited financial indicators when discussing the health of Social Security: the *cost rate*, the *trust fund balance*, and the *trust fund ratio*. The cost rate equals the overall cost of the Social Security program in a given year divided by the taxable payroll for that year. The trust fund balance equals the difference between projected annual income and projected annual cost, as a percentage of the taxable payroll. The trust fund ratio equals the assets of the Social Security Trust Funds at the beginning of a calendar year divided by the expected expenditure for that year.

We collect all forecasts for each measure published in the annual Trustees Reports from 1978, when the reports began consistent reporting of financial indicators, until 2013. The reports usually include yearly forecasts between the year of the report and 10 years in the future and then every fifth subsequent year. After 2000, single-year supplemental tables are available online.

These three financial indicators directly relate to the economic and public policy debates that have occurred over nearly the entire lifetime of Social Security. After the Social Security Amendments of 1983, for example, the trust fund balance increased primarily because of higher payroll tax rates intended to build up a surplus in the trust fund, although benefit levels increased, too. Numerous economic studies find Social Security affects personal savings through reduction of disposable income because of payroll taxes and anticipated benefits during retirement (Harris 1941; Feldstein 1974; Diamond and Hausman 1984). Gramlich (1996) argued that proposed Social Security reform faces competing challenges in political economy: ensuring long-run actuarial balance while not lowering the ratio of discounted benefits to discounted taxes paid (the “money’s worth” ratio). The long-run actuarial balance, a function of the trust fund balance and cost rate, can

Figure 4
Cost of Mortality Forecasting Errors



Notes: Each panel of the figure corresponds to a Trustees Report. Within each panel, we plot the forecast error in total Social Security expenditures (solid lines) and the forecast error in total Social Security expenditures due to mortality forecasting errors (dashed lines). Finally, we represent the Great Recession as a vertical grey region.

be maintained by raising payroll tax rates or lowering benefit levels, although such changes would reduce the money's worth ratio. Other proposed reforms, such as individual accounts and personal savings accounts, offer a possibility of extending the solvency of Social Security and maintaining the money's worth ratio but face intense public scrutiny (Samwick 1999).

The Cost of Mortality Forecasting Errors

Before turning to the three financial indicators, we begin by comparing the forecast errors in cost that are specifically due to forecast errors in mortality (dashed line, Figure 4) to the overall forecast errors in cost (solid line). In theory, either of these forecast errors in cost could be larger than the other because forecast errors in cost potentially come from many inputs other than mortality.

For each Trustees Report and forecast year, we estimate the number of additional retirees that the Social Security Administration did not expect because of errors in predicting life expectancy. For example, the 2005 Trustees Report underforecasted male life expectancy at age 65 in the year 2010 by 1.3 years (forecast: 16.6 years; true outcome: 17.9 years). The 1.3 years under-forecast of life expectancy corresponds to approximately 151,000 male beneficiaries. We estimate the forecast errors in costs due to forecast error in mortality as the product of the total number of additional beneficiaries and the average benefit amount per year. For this figure, we plot the forecast year on the horizontal axis and the forecast error in cost (in billions of 2010 dollars) on the vertical axis. Each panel presents forecasts from different Trustees Reports. The years of the Great Recession are denoted by the grey shaded region. To put the figure into perspective, the total cost of the Social Security program in 2010 was $712.5 billion.

Figure 4 emphasizes four points. First, mortality is a highly predictable part of the overall forecast error in cost, as evidenced by the highly smooth and almost linear dashed lines in each panel. Second, for many years, forecast errors in cost specifically due to forecast errors in mortality were a large fraction of the overall forecast error in cost. Third, the forecast errors in cost due to forecast errors in mortality are neither random nor constant. The errors increase secularly and thus strongly suggest the existence of information that can be used to improve forecasting performance. Finally, the overall forecast errors in costs are highly variable relative to errors due to mortality. They are much larger during the Great Recession, shown by the vertical shaded area, but these overall forecast errors in costs were also large at times well before the onset of the Great Recession.

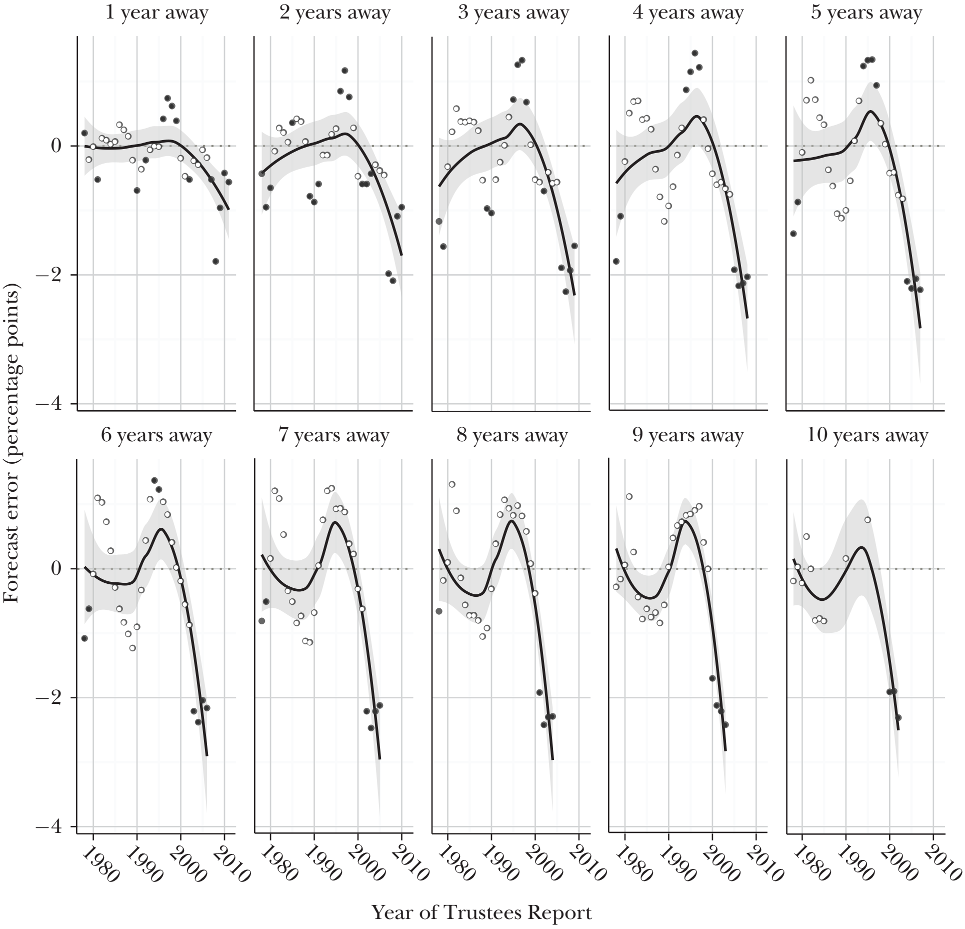
Cost Rate Forecasting Errors

Figure 5 reports the forecast error in the cost rate (the vertical axis in each panel) made in a Trustees Report in the given year (the horizontal axis in each panel) for a number of years out into the future (in the title of each panel). For example, the upper-left panel shows the forecast error in the cost rate for forecasts made one year in advance of the year forecast. A value of zero on the vertical axis indicates that the forecast was perfectly accurate. White points fall within the forecast uncertainty interval and black points fall outside. To enhance readability, we superimpose on each panel a smoothed line showing the path of the errors.[4]

The pattern of forecast errors in Figure 5 is striking. Forecasts from Trustees Reports until about 2000 were approximately unbiased, which can be seen by the roughly random scatter of points vertically around the horizontal line at zero forecast error. However, forecasts from Trustees Reports after roughly the year 2000 were increasingly biased over time, and all in the same direction. Congress and

[4] The smoothed line is estimated with a locally weighted scatterplot smoothing (LOESS) procedure, in which the predicted error *t* for Trustees Report is calculated based on a local polynomial of degree 2, fit to neighboring observations. These observations are weighted by their tricubic distance from the Trustees Report in question.

Figure 5
Cost Rate Forecasting Errors

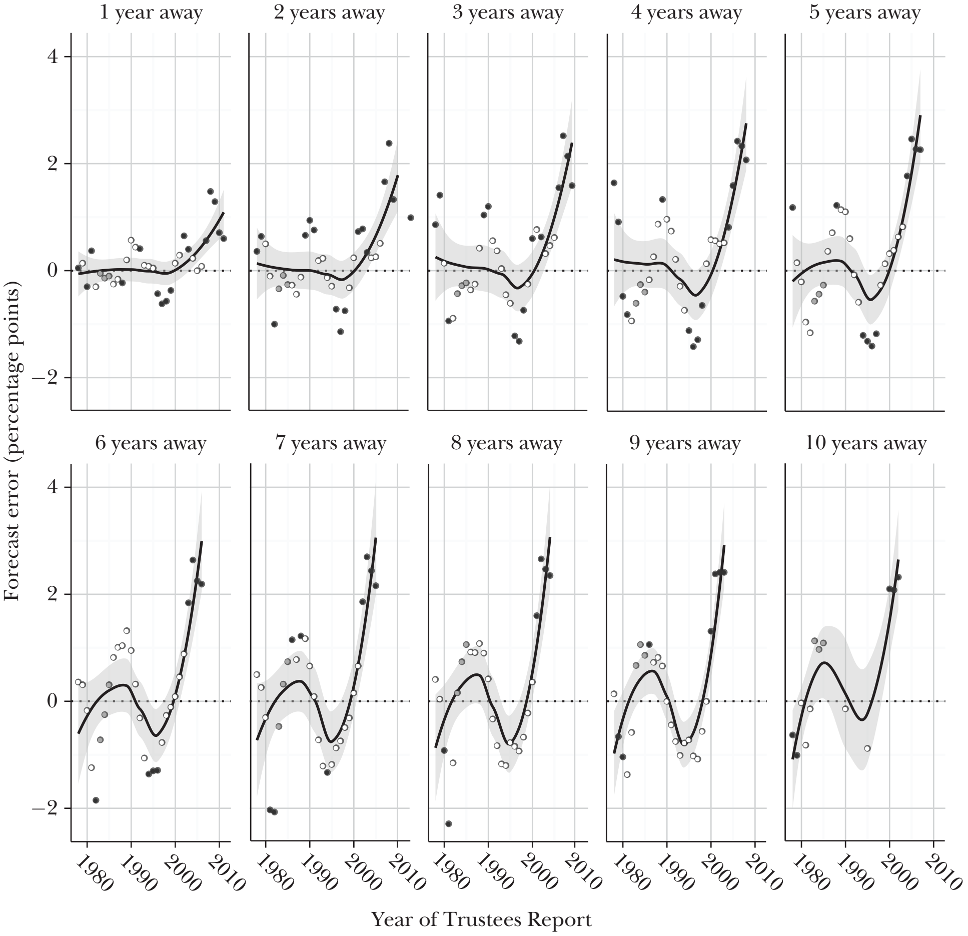


Notes: The graphs show forecast errors in the cost rate (vertically) by the year of the forecast (horizontally) by how many years into the future the forecast is made (in the title of each panel). Cost rate forecasting errors are overestimates if positive and underestimates if negative. Points are white if the error is within the Social Security Administration's uncertainty interval and black otherwise. To enhance readability, we superimpose on each panel a smoothed line showing the path of the errors. This error bar around this line relates to the observed path of the errors and not to the SSA predicted path of observations. See footnote 4.

other users of these forecasts would have been misled into thinking that the cost of the Social Security program was less than it actually turned out to be. This is as true for forecasts one year into the future (top left) as for forecasts 10 years into the future (bottom right). As expected, the errors are larger for forecasts farther into the future.

Finally, Figure 5 shows that the largest errors are also most likely to be outside the uncertainty intervals (as indicated by black dots). The purpose of uncertainty

Figure 6
Trust Fund Balance Forecasting Errors



Notes: The graphs show forecast errors in balance (vertically) by the year of the forecast (horizontally) by how many years into the future the forecast is made (in the title of each panel). Positive errors overestimate the Trust Fund balance; negative errors underestimate it. Points are white if the error is within the Social Security Administration's uncertainty interval and black otherwise. To enhance readability, we superimpose on each panel a smoothed line showing the path of the errors. This error bar around this line relates to the path of the observed errors and not to the SSA predicted path of observations. See footnote 4.

estimates is to protect oneself from drawing overconfident conclusions from the data, and if estimates are consistently falling outside those uncertainty intervals, then alteration and improvement in the forecasting process should follow.

Trust Fund Balance Forecasting Errors

A positive annual trust fund balance indicates the program has a surplus for the year and a negative trust fund balance translates to a deficit. We present Figure 6 in the same format as Figure 5. The evaluation of forecasting errors in the trust fund balance leads us to the same conclusions as forecasting errors in the cost rate. The Social Security Administration forecasts of trust fund balances were approximately unbiased until about 2000, after which they become substantially biased. Moreover, the direction of the biases are all in the same direction, making the Social Security trust funds look healthier than they turned out to be. The reported uncertainty intervals are again overconfident.

Trust Fund Ratio Forecasting Errors

When the trust fund ratio equals 0 percent or becomes negative, the Social Security Trust Funds are insolvent. The Trust Funds are deemed financially adequate in the short term if the ratio stays above 100 percent for the first 10 forecasted years. Insolvency does not release the federal government from its obligation to pay some level of benefits to qualified individuals (Meyerson 2014). The Social Security Act stipulates that every fully insured individual is entitled to receive benefits. On the other side, the Antideficiency Act prohibits the federal government from paying Social Security benefits beyond the balance of the Trust Funds. Once insolvency occurs, beneficiaries would either receive delayed or lower benefit payments.

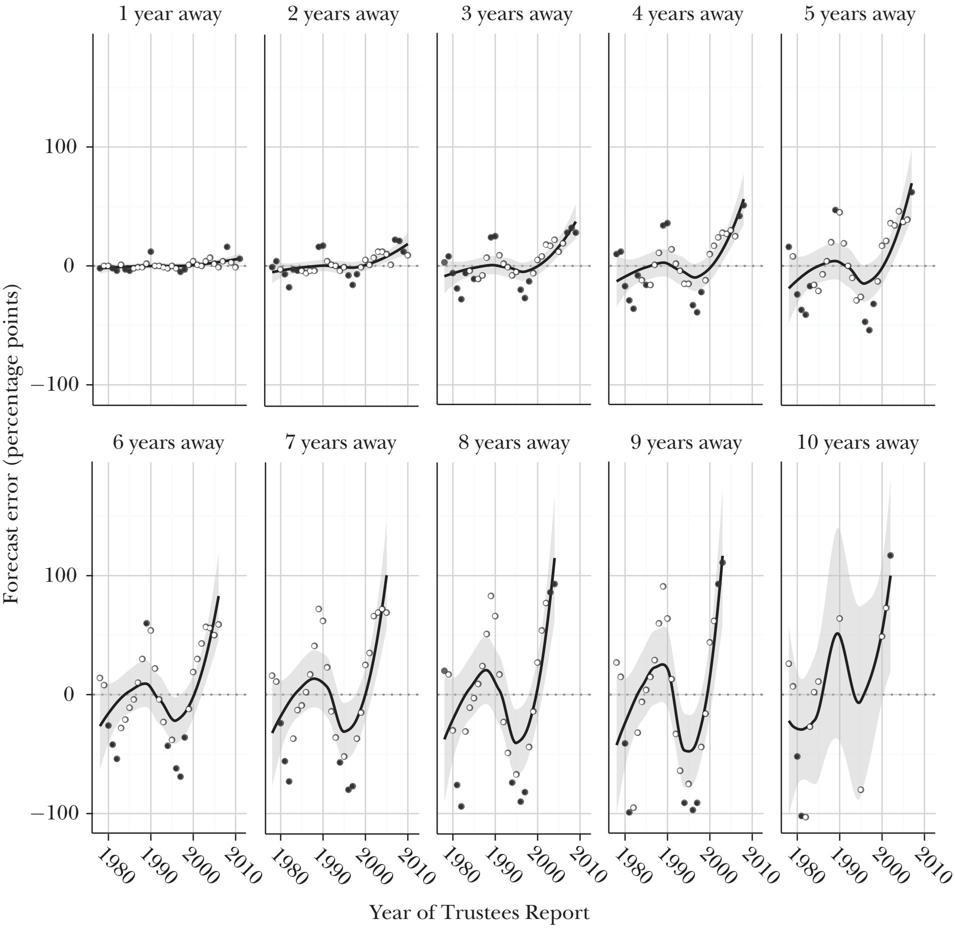
In Figure 7, we present results in a form parallel to Figures 5 and 6. While the uncertainty intervals appear to have better coverage when compared to the cost rate and trust fund balance metrics, the overall results in this figure confirm the main results from our analysis of the cost rate and trust fund balance. First, trust fund ratio forecast errors are approximately unbiased from 1978 through about the year 2000, as indicated by the dots scattered randomly above and below the vertical line drawn at zero. After 2000, forecast errors became increasingly biased, and in the same direction. Trustees Reports after 2000 all overestimated the assets in the program and overestimated solvency of the Trust Funds. The size of this bias has increased over time, with the more recent Trustee Reports being less and less reliable. Finally, the coverage of uncertainty estimates did not improve over time and were strongly and positively correlated with the size of the absolute error.

Implication of Financial Forecasting Errors for Proposal Scoring

In addition to producing the annual Trustees Report, the Office of the Chief Actuary also scores policy proposals to alter Social Security submitted by members of Congress, the administration, and select professional organizations. For each of the policy proposals it scores, the Office of the Chief Actuary makes point-estimate predictions about what would happen to one or more financial metrics, such as those we study above, if the proposal became law. Although the Office of the Chief Actuary includes no uncertainty measures with these predictions, we can estimate their uncertainty on the basis of our evaluation of their forecasts.

The extent of uncertainty in these counterfactual predictions can be divided into two components. The first is the inherent uncertainty of the effect of the intervention if the law changes as proposed. The second is the uncertainty in forecasting

Figure 7
Trust Fund Ratio Forecasting Errors



Notes: The graphs show forecast errors in the trust fund ratio (vertically) by the year of the Trustees Report forecast (horizontally) by how many years into the future the forecast is made (in the title of each panel). Positive errors overestimate the Trust Fund ratio; negative errors underestimate it. Points are white if the error is within the Social Security Administration's uncertainty interval and black otherwise. To enhance readability, we superimpose on each panel a smoothed line showing the path of the errors. This error bar around this line relates to the path of the observed errors and not to the SSA predicted path of observations. See footnote 4.

the same financial indicators under current law, as we do earlier. We use our evaluation of the second component as a lower bound for the uncertainty of the Office of the Chief Actuary's policy scoring.

The Social Security Administration evaluated 93 proposals since 2000, which resulted in 110 assessments of financial indicators for which we can evaluate forecasting performance (see http://ssa.gov/oact/solvency). For example, in 2015, the Social Security Administration evaluated the effect of President Obama's Executive Actions for immigration on Social Security solvency. The Chief Actuary concluded immigration reform would increase the cost rate by an average 0.04 percent over the next 75 years, which is considerably smaller than most cost rate forecasting errors made since 2000. Overall, we found that 42 percent of policy assessments by the Office of the Chief Actuary predicted changes in Social Security finances that were smaller than the average forecasting error made since 2000. And 95 percent of the assessments concluded by predicting changes in Social Security finances that were smaller than the maximum forecasting error made since 2000.

Members of Congress and the public devote considerable energy debating policy proposals on the basis of these evaluations. Presidents and their opponents tout the merits of policy proposals to engender public support. But if the lower bound on the magnitude of forecasting errors exceeds the estimated effect of the reforms, then it seems likely these discussions and debates are not grounded in the best information available.

Conclusions and Recommendations

In recent years, especially after about 2000, the Social Security Administration began issuing systematically biased forecasts with overconfident assessments of uncertainty. Reliance on such forecasts led policymakers and other users of the forecasts to conclude that the Social Security Trust Funds were on firmer financial ground than actually turned out to be the case. We focus here on three steps that the Social Security Administration should take to ensure this problem is addressed; other suggestions are offered in King, Kashin, and Soneji (2015a).

First, forecasting mistakes are no embarrassment unless the forecaster fails to learn from them. Thus, we recommend that the Social Security Administration publish annually a systematic and comprehensive evaluation of its forecasting performance for both demographic factors and financial solvency. This best practice of forecasting self-evaluation is routine among academic researchers (Lee and Miller 2001) and professional actuaries (Lu and Won 2011), for social security programs in other countries (Shaw 2007), and in other parts of the US government, such as the Congressional Budget Office (2013), the Census Bureau (Wang 2002), the Bureau of Labor Statistics (Wyatt 2010), and even other parts of the Social Security Administration itself (US Social Security Administration's "Fiscal Year 2013 Major Evaluations" and *Social Security Administration Agency Financial Report: Fiscal Year* 2013). Every future Trustees Report, without exception, should include a routine evaluation of all prior forecasts, and a discussion of what forecasting mistakes were made, what was learned from the mistakes, and what actions might be taken to improve forecasts going forward.

Second, the Social Security Administration withholds from public view much of the data and procedures it uses to make many of its forecasts. The Office of the

Chief Actuary, which produces the demographic and economic forecasts, does not share much of its data and procedures even with other parts of the Social Security Administration. Currently, the best anyone can do to understand how the Social Security Administration forecasts work is to attempt to reverse-engineer their results (as done by many involved in the policy process and authors of simulation programs such as SSASIM by the Policy Simulation Group; see also King and Soneji 2011, and Soneji and King 2012). The "replication standard" for data sharing is the widely understood and accepted best practice throughout the scientific community (King 1995, p. 444) and echoed in the Obama administration's executive orders requiring "a presumption in favor of openness," and that data produced by government be "accessible, discoverable, and usable by the public" (Executive Office of the President, Memorandum, May 9, 2013).

Finally, it appears to us and to other outside observers that the forecasting procedures used by the Social Security Administration are in a number of ways ad hoc and suboptimal. These approaches also fail to take advantage of many of the dramatic improvements in statistical modeling over the last several decades (for example, Girosi and King 2008; King and Soneji 2011). Even some explicitly quantitative parts of the methods seem idiosyncratic or unnecessarily model dependent.

Our study reveals systematic errors in both demographic and solvency forecasts. Forecasting errors in economic variables, such as the labor force participation rate and growth in average wages, may also contribute to systematic errors in Trust Fund solvency forecasts. For the disability program of Social Security, forecasting errors in the disability incidence rate may be an additional important source of solvency forecast error.

This list of "best practices" is neither new nor controversial. There is a Social Security Advisory Board Technical Panels on Assumptions and Methods made up of outside experts. The Social Security Administration's own outside advisors have repeatedly and emphatically recommended that the Office of the Chief Actuary make its data and procedures widely available, and allow its work to be replicated by outside groups. The collective efforts of the scientific community could easily be marshaled to improve the difficult forecasting task that confronts the Social Security Administration, all essentially without cost to the taxpayer. The creation of transparent forecasting procedures will also enable members of Congress and partisans on all sides to consider alternative assumptions explicitly when they debate proposals to ensure the solvency of Social Security. Forecasts of Social Security solvency also shape debates on immigration, public health, taxation, and income redistribution from working age adults to retirees. Accurate forecasts would help ensure these debates are based on the best information available.

■ *For helpful advice or comments, we are grateful to Bill Alpert, Jim Alt, Steve Ansolabehere, Neal Beck, Nicholas Christakis, Mo Fiorina, Dan Gilbert, Alexander Hertel-Fernandez, Martin Holmer, David Langer, and Theda Skocpol. We received approval from our university Institutional Review Boards for this study.*

References

Congressional Budget Office, US Congress. 2013. "CBO's Economic Forecasting Record: 2013 Update." January.

Diamond, Peter, and Jerry Hausman. 1984. "Individual Retirement and Savings Behavior." *Journal of Public Economics* 23(1–2): 81–114.

Diamond, Peter A., and Peter R. Orszag. 2005. *Saving Social Security: A Balanced Approach.* Revised edition (paperback). Brookings Institution Press.

Executive Office of the President, Office of Management and Budget. 2013. "Memorandum for the Heads of Executive Departments and Agencies." M-13-13. May 9. https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

Feldstein, Martin. 1974. "Social Security, Induced Retirement, and Aggregate Capital Accumulation." *Journal of Political Economy* 82(5): 905–26.

Girosi, Federico, and Gary King. 2008. *Demographic Forecasting.* Princeton University Press.

Gramlich, Edward M. 1996. "Different Approaches for Dealing with Social Security." *Journal of Economic Perspectives* 10(3): 55–66.

Granados, José A. Tapia. 2005. "Increasing Mortality during the Expansions of the US Economy, 1900–1996." *International Journal of Epidemiology* 34(6): 1194–1202.

Harris, Seymour E. 1941. *Economics of Social Security.* New York: McGraw-Hill.

Human Mortality Database. N.d. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org (data downloaded on June 4, 2014).

Kashin, Konstantin, Gary King, and Samir Soneji. 2015a. "Explaining Nontransparency and Increasing Bias in U.S. Social Security Administration Forecasts." http://j.mp/SSAbias. (Forthcoming in *Political Analysis.*)

Kashin, Konstantin, Gary King, and Samir Soneji. 2015b. "Replication Data for: Systematic Bias and Nontransparency in US Social Security Administration Forecasts." UNF:5:1oerGFXQ0Bu9bcMFU5/t2A== http://dx.doi.org/10.7910/DVN/28122. Harvard Dataverse [Distributor] V1 [Version].

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28(3): 443–99. http://j.mp/jCyfF1.

King, Gary, and Samir Soneji. 2011. "The Future of Death in America." *Demographic Research* 25(1): 1–38. http://j.mp/iXUpBv.

Lee, Ronald, and Timothy Miller. 2001. "Evaluating the Performance of the Lee–Carter Method for Forecasting Mortality." *Demography* 38(4): 537–49.

Lu, Joseph, and Wun Wong. 2011. "Mortality Improvement in the USA: Analysis, Projections and Extreme Scenarios." Technical report, Society of Actuaries Schaumburg, IL.

Manski, Charles F. 2013. *Public Policy in an Uncertain World: Analysis and Decisions.* Harvard University Press.

Meyerson, Noah P. 2014. *Social Security: What Would Happen If the Trust Funds Ran Out?* Technical report. Federal Publications, Key WorkPlace Documents at DigitalCommons@ILR.

Ruhm, Christopher J. 2000. "Are Recessions Good for Your Health?" *Quarterly Journal of Economics* 115(2): 617–650.

Samwick, Andrew A. 1999. "Social Security Reform in the United States." *National Tax Journal* 52(4): 819–842.

Shaw, Chris. 2007. "Fifty Years of United Kingdom National Population Projections: How Accurate Have They Been?" *Population Trends* 128: 8–23.

Soneji, Samir, and Gary King. 2012. "Statistical Security for Social Security." *Demography* 49(3): 1037–1060. http://j.mp/Qvla7N.

Stuckler, David, Christopher Meissner, Price Fishback, Sanjay Basu, and Martin McKee. 2011. "Banking Crises and Mortality during the Great Depression: Evidence from US Urban Populations, 1929–1937." *Journal of Epidemiology and Community Health* 66(5): 410–19.

Wang, Ching-li. 2002. "Evaluation of Census Bureau's 1995–2025 State Population Projections." Working Paper No. POP-WP067, US Census Bureau.

Wyatt, Ian D. 2010. "Evaluating the 1996–2006 Employment Projections." *Monthly Labor Review,* September.

US Social Security Administration. 2012. "Fiscal Year 2013 Major Evaluations." Charts. http://j.mp/SSAevals.

US Social Security Administration. 2013. *Social Security Administration Agency Financial Report: Fiscal Year 2013.*

US Social Security Administration. Various Years. *Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds.*

Journal of Economic Perspectives—Volume 29, Number 2—Spring 2015—Pages 259–266

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., Saint Paul, Minnesota, 55105.

Smorgasbord

Education Next has published a group of articles in its Spring 2015 issue on "Revisiting the Moynihan Report on its 50th Anniversary." As one example, Sara McLanahan and Christopher Jencks contribute "Was Moynihan Right?" "Moynihan's claim that growing up in a fatherless family reduced a child's chances of educational and economic success was furiously denounced when the report appeared in 1965, with many critics calling Moynihan a racist. For the next two decades few scholars chose to investigate the effects of father absence, lest they too be demonized if their findings supported Moynihan's argument. Fortunately, America's best-known black sociologist, William Julius Wilson, broke this taboo in 1987, providing a candid

■ *Timothy Taylor is Managing Editor,* Journal of Economic Perspectives, *based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

http://dx.doi.org/10.1257/jep.29.2.259 doi=10.1257/jep.29.2.259

assessment of the black family and its problems in *The Truly Disadvantaged*. Since then, social scientists have accumulated a lot more evidence on the effects of family structure." As an example, they cite the literature review by Sara McLanahan, Laura Tach, and Daniel Schneider concerning "The Causal Effects of Father Absence" in the *Annual Review of Sociology* (July 2013, pp. 399–427). These authors conclude: "We find strong evidence that father absence negatively affects children's social-emotional development, particularly by increasing externalizing behavior. These effects may be more pronounced if father absence occurs during early childhood than during middle childhood, and they may be more pronounced for boys than for girls. There is weaker evidence of an effect of father absence on children's cognitive ability." The *Education Next* articles are at http://educationnext.org/revisiting-moynihan-report-50th-anniversary. The McLanahan, Tach, and Schneider article appears in the *Annual Review of Sociology*, July 2013, pp. 399–427, and is at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3904543.

An all-star list of environmental and welfare economists discuss "Should Governments Use a Declining Discount Rate in Project Analysis?" Kenneth J. Arrow, Maureen L. Croppery, Christian Gollierz, Ben Groom, Geoffrey M. Heal, Richard G. Newell, William D. Nordhaus, Robert S. Pindyck, William A. Pizer, Paul R. Portney, Thomas Sterner, Richard S. J. Tol, and Martin L. Weitzman write: "In the United States, however, the Office of Management and Budget (OMB) recommends that project costs and benefits be discounted at a constant exponential rate (which, other things equal, assigns a lower weight to future benefits and costs than a declining rate), although a lower constant rate may be used for projects that affect future generations. . . . For intragenerational projects, the OMB (2003) recommends that benefit-cost analyses be performed using a discount rate of 7 percent, representing the pretax real return on private investments, and also a discount rate of 3 percent, representing the "social rate of time preference. . . . France and the United Kingdom use discount rate schedules in which the discount rate applied today to benefits and costs occurring in the future declines over time. That is, the rate used today to discount benefits from year 200 to year 100 is lower than the rate used to discount benefits in year 100 to the present." "We have argued that theory provides compelling arguments for using a declining certainty equivalent discount rate. . . . Establishing a procedure for estimating a DDR for project analysis would be an improvement over the OMB's current practice of recommending fixed discount rates that are rarely updated." In *Review of Environmental Economics and Policy*, Summer 2014, vol. 8, no. 2, pp. 145–163. http://reep.oxfordjournals.org/content/8/2/145.full.pdf+html.

Jonathan V. Hall and Alan B. Krueger offer "An Analysis of the Labor Market for Uber's Driver-Partners in the United States." "From a base of near zero in mid-2012, more than 160,000 drivers actively partnered with Uber at the end of 2014 in the United States, and the rate of growth was rising throughout this period. In the United States driver-partners received $656.8 million in payments from Uber during the last three months of 2014." "[A]lthough some have argued that the sharing economy is weakening worker bargaining power and responsible for much of the rise in inequality in the United States, the actual effect is much more complicated and less clear. First, there is little evidence of a secular rise in the percentage of workers who are self-employed, independent contractors, or part-time. . . . Second, inequality increased dramatically in the United States long before the advent of the sharing economy, and has increased much less in many other countries that, unlike the United States, experienced a sharp rise in part-time work. Third, at least insofar as the advent of ride sharing services like Uber is concerned, the relevant market comparison is to other for-hire drivers, many of whom were independent contractors prior to the launch of Uber. Moreover, the availability of modern technology, like the Uber app, provides many advantages and lower prices for consumers compared with the traditional taxi cab dispatch system, and this has boosted demand for ride services, which, in turn, has increased total demand for workers with the requisite skills to work as for-hire drivers, potentially raising earnings for all workers with such skills. And finally, the growth of Uber has provided new opportunities for driver-partners, who . . . seem quite pleased to have the option available." Princeton University, Industrial Relations Section, Working Paper #587, January 22, 2015. http://dataspace.princeton.edu/jspui/bitstream/88435/dsp010z708z67d/5/587.pdf.

Elroy Dimson, Paul Marsh, and Mike Staunton ask: "Responsible Investing: Does It Pay to be Bad?" "The paradox, then, is that depressed share prices for what some regard as noxious and nasty businesses may demonstrate that responsible and ethical investors are having an impact on the value of a company whose activities conflict with social norms. If so, the shares will ultimately sell at a lower price relative to fundamentals. For example, they may trade at a lower price/earnings or lower price/dividend ratio. Buying them would then offer a superior expected financial return which, for some investors, compensates for the emotional 'cost' of exposure to offensive companies." "[I]f companies have a lower stock price, they offer a buying opportunity to investors who are relatively untroubled by ethical considerations." "To maximize the probability of success as an activist, asset owners might consider the 'washing machine' strategy . . . [A] large investor can generate continuing outperformance by buying non-responsible companies and turning them into more responsible businesses. After they have been cleaned up, the shares may then be sold at a price that reflects the accomplishments of the activist." *Credit Suisse Global Investment Returns Yearbook 2015*, February 2015, pp. 17–27. At http://publications.credit-suisse.com/tasks/render/file/index.cfm?fileid=AE924F44-E396-A4E5-11E63B09CFE37CCB.

Jan Luiten van Zanden, Joerg Baten, Marco Mira d'Ercole, Auke Rijpma, Conal Smith, and Marcel Timmer have edited a collection of 13 chapters in the book, *How Was Life? Global Well-being Since 1820.* "How was life in 1820, and how has it improved over time? . . . The report examines 10 individual dimensions of well-being, tracking them over time and space, then pulls them together in a new composite indicator. The dimensions covered reflect a broad range of material and non-material aspects of well-being: per capita GDP, real wages, educational attainment, life expectancy, height, personal security, political institutions, environmental quality, income inequality and gender inequality. . . . For some of these dimensions the statistical correlation with the evolution of GDP per capita is strong. Education (as measured by literacy and educational attainment) and health status (as measured by life expectancy and height) improved strongly in many countries in the world, and there is a strong cross-section and over-time correlation with GDP per capita. . . . The statistical correlation with GDP per capita was much weaker for other well-being dimensions. Political institutions (as measured by electoral participation and competition) have greatly improved around the world in the past century. But their development was far from gradual, with sometimes violent swings in political rights in some countries. Also differences across countries in personal security (measured by homicide rates and exposure to conflict) do not correlate well with GDP per capita. . . . A negative correlation with GDP per capita is clearly in place when looking at quality of the environment. . . . Long-term trends in income inequality, as measured by the distribution of pre-tax household income across individuals, followed a U-shape in most Western European countries and Western Offshoots. It declined between the end of the 19th century until about 1970, followed by a rise. . . . Gender inequality as measured by outcomes in health status, socio-economic status and political rights, has been on a declining trend over the past 60 years in most world regions. . . . The Composite indicator of well-being presented in this report indicates that progress in well-being has been widespread since the early 20th century, with the possible exception of Sub-Saharan Africa. The evidence presented in this report also suggests that since the 1970s between-country inequality in composite well-being has been lower than in GDP per capita, while being more pronounced in the period before." OECD, 2014. It can be read at http://www.oecd-ilibrary.org/economics/how-was-life_9789264214262-en.

Matthew Rognlie takes on the task of "Deciphering the Fall and Rise in the Net Capital Share." "The story of the postwar net capital share is not a simple one. It has fallen and then recovered—with a large long-term increase in net capital income from housing . . . Given the important role of housing, observers concerned about the distribution of income should keep an eye on housing costs . . . Beyond housing, the results in this paper (if anything) tentatively suggest that concern about inequality should be shifted away from the split between capital and labor, and toward other aspects of distribution, such as the within-labor distribution of income." *Brookings Papers on Economic Activity*, Spring 2015. http://www.brookings.edu/about/projects/bpea/papers/2015/land-prices-evolution-capitals-share.

Paul H. Rubin delivered the 2013 Presidential Address to the Southern Economic Association on the subject of "Emporiophobia (Fear of Markets): Cooperation or Competition?" "In their economic lives, people produce goods and services and exchange these goods and services for others. Both the production of goods and the exchange of goods for other goods are cooperative acts. There is no competition in these actions. The motive for some acts may be competitive, but the actions themselves are cooperative. . . . Unless an agent is willing to engage in illegal actions (for example, burning a competitor's factory) or willing to go outside the market (e.g., complaining to the Federal Trade Commission about a competitor), any competitive act is actually performed through cooperative behavior. . . . Adam Smith is the father of competitive analysis. But he is also the father of cooperative analysis. Specialization is the mother of cooperation. The pin factory is a masterful analysis of cooperation. Somehow we economists have made the competitive analysis in Smith the basis for our discipline and have made cooperation into something of a stepchild. . . . Wal-Mart succeeds not because it has beat up its rivals and driven them out of business. It succeeds because it has done a better job of cooperating with consumers, by offering them stuff they want at the lowest possible prices. Of course, economists know this, but since non-economists begin with the competition model, economists must be defensive and try to dissuade citizens of their prior beliefs. If the default way of thinking was cooperation, then the critics of markets would be on the defensive." *Southern Economic Journal* April 2014, vol. 80, no. 4, pp. 875–89. Also available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2360674.

Productivity Growth

Antonin Bergeaud, Gilbert Cette, and Rémy Lecat discuss "Productivity Trends from 1890 to 2012 in Advanced Countries." "We can mainly distinguish four periods from 1890 to 2012. . . . 1. From 1890 to WWI, productivity was growing moderately and was characterized by a UK leadership and a catch-up by the other countries. 2. After the WWI slump, the Interwar and WWII years were characterized by a heightening of the US leadership, as it experienced an impressive big wave of productivity acceleration in the 1930s and 1940s, while other countries struggled with the Great Depression legacy and WWII. 3. After WWII, European countries and Japan benefited from the big wave experienced earlier in the United States. 4. Since 1995, the post-war convergence process has come to an end as US productivity growth overtook Japan and other countries', although it is not up to its 1930s or 1940s pace. Shorter and smaller than the first one, a second big wave appeared in the US and, in a less explicit way, in the other areas. Bank of France, Working Paper 475, February 2014. https://www.banque-france.fr/uploads/tx_bdfdocumentstravail/DT_475.pdf.

The McKinsey Global Institute tackles the question: "Global Growth: Can Productivity Save the Day in an Aging World?" For low and middle-income countries: "Overall, it is striking that the absolute gap between productivity in emerging and developed economies has not narrowed. Productivity in developed economies today remains almost five times that of emerging economies. Narrowing this gap is one of the biggest opportunities for—and challenges to—long-term global growth." For high-income countries: "In developed economies, more than half—55 percent—of the productivity gains that MGI's analysis finds are feasible could come from closing the gap between low-productivity companies and plants and those that have high productivity. There are opportunities to continue to incorporate leaner supply-chain operations throughout retail, and to improve the allocation of the time spent by nurses and doctors in hospitals and health-care centers, for example. Across countries, large differences in average productivity within the same industry indicate industry-wide opportunities for improvement. For instance, low productivity in retail and other service sectors in Japan and South Korea reflects a large share of traditional small-scale retailers. High costs in the US health-care system partly reflect the excessive use of clinically ineffective procedures. Even agriculture, automotive manufacturing, and other sectors that have historically made strong contributions to productivity growth have ample room to continue to diffuse innovations and become more efficient." January 2015. At: http://www.mckinsey.com/insights/growth/can_long-term_global_growth_be_saved.

Islamic Banking

The *World Islamic Banking Competitiveness Report 2014–15: Participation Banking 2.0* notes that what used to be called "Islamic banking" is now sometimes called "Participation banking." "The global profit pool of Participation banks is set to triple over the next five years. For the first time in history, the combined profit of Participation banks crossed $10 bn (2013). . . . But this is only half the story. The ROEs [return on equity] of Participation banks remain c. 19% lower when compared to traditional banks in the same markets. . . . This year, EY studied 2.2 million customer sentiments dispersed across countless on-line sources in nine key markets (Saudi Arabia, Bahrain, Kuwait, the UAE, Malaysia, Indonesia, Turkey, Qatar, and Oman). Results show that for many Participation banks, customer satisfaction is, at best, mediocre." EY, http://www.ey.com/EM/en/Industries/Financial-Services/Banking---Capital-Markets/EY-world-islamic-banking-competitiveness-report-2014-15.

Renee Haltom discusses "Islamic Banking, American Regulation: For Some American Muslims, Sharia-Compliant Banks Are an Important Part of the Financial Landscape." "According to research by Feisal Khan, an economics professor at Hobart and William Smith Colleges in upstate New York, most Islamic finance transactions are economically indistinguishable from traditional, debt- and interest-based finance. Where there is principal and a payment plan, there is an implied interest rate, Khan argued . . . He is not the first economist to make such a claim. Many Islamic scholars argue that murabaha contracts don't share risk and thus are not Sharia-compliant—and experts estimate that such contracts constitute up to 80 percent of the global Islamic finance volume. Other economists have noted that the terms of Islamic financial contracts often move with market interest rates. In the United States, Islamic financial products are frequently marketed with information about implied interest rates to allow customers to compare prices or simply to comply with American regulation. . . . In 1997, the United Bank of Kuwait (UBK), which then had a branch in New York, requested interpretive letters from its regulator, the Office of the Comptroller of the Currency (OCC), on ijara and murabaha mortgage products. The OCC approved them on the very grounds that they were economically equivalent to traditional products. . . . Possibly because the products are unfamiliar to many investors, there is a smaller secondary market for Islamic financial products, so it has been harder for Islamic mortgage lenders to remain liquid, hindering the market's growth. In the United States, housing agencies Freddie Mac and Fannie Mae started buying Islamic mortgage products in 2001 and 2003, respectively, to provide liquidity, and they are now the primary investors in Islamic mortgages." *Econ Focus*, Federal Reserve Bank of Richmond, Second Quarter 2014, pp. 15–19.

Discussion Starters

The Copenhagen Consensus Center has been commissioning research to identify what development proposals would have the highest benefit-cost ratios. Three prominent economists—Finn Kydland, Tom Schelling, and Nancy Stokey—have evaluated that research and recommended 19 areas of focus, all of which promise benefits at least 15 times greater than costs. Their list includes: 1) Lower chronic child malnutrition by 40%. 2) Halve malaria infection. 3) Reduce tuberculosis deaths by 90%. 4) Avoid 1.1 million HIV infections through circumcision. 5) Cut early death from chronic disease by 1/3. 6) Reduce newborn mortality by 70%. 7) Increase immunization to reduce child deaths by 25%. 8) Make family planning available to everyone. 9) Eliminate violence against women and girls. 10) Phase out fossil fuel subsidies. 11) Halve coral reef loss. 12) Tax pollution damage from energy. 13) Cut indoor air pollution by 20%. 14) Reduce trade restrictions. 15) Improve gender equality in ownership, business and politics. 16) Boost agricultural yield growth by 40%. 17) Increase girls' education by two years. 18) Achieve universal primary education in sub-Saharan Africa. 19) Triple preschool in sub-Saharan Africa. The website of the Copenhagen Consensus Center is at http://www.copenhagenconsensus.com. Over 100 peer-reviewed analyses of different proposals are available at http://www.copenhagenconsensus.com/post-2015-consensus/research. The list of 19 goals from Kydland, Schelling, and Stokey is at http://www.copenhagenconsensus.com/post-2015-consensus/nobel-laureates-guide-smarter-global-targets-2030-0.

Eugene Gholz tells the story of a seeming crisis and how it was largely resolved by market forces in "Rare Earth Elements and National Security." "By the early 2000s China produced 97 percent of the world's REEs. Because of REEs' extreme supplier concentration and the wide acceptance that these materials are vital inputs to military products, concerns over potential supply-chain vulnerabilities soon began to percolate around the developed world. . . . U.S. government agencies, led by the Department of Defense and the U.S. Geological Survey, quietly began to study the risks of dependency on China . . . In early September 2010, in the midst of a maritime border dispute, Japan detained the captain of a Chinese fishing trawler. Afterward, China allegedly embargoed exports of rare earth oxides, salts, and metals to Japan. (Japanese companies insisted the embargo was real, even as the Chinese government officially denied it.) . . . Prices soared in the REE spot market in the wake of China's 2010 export cuts, especially as downstream users—companies that incorporate REEs into other products—filled inventories to protect themselves from future disruptions." But the market responded with supply increases and substitutes for demand. “Motivated by expected increases in demand, investors in the United States, Japan, and Australia were already opening rare earth mines and building new processing capabilities by 2010, and other investors were moving ahead on mines around the world in places like Canada, South Africa, and Kazakhstan. . . . When rare earth prices surged in 2010, even more potential entrants swarmed. . . . [A]t the time of China’s 2010 export embargo to Japan, the largest-volume use of rare earths was in gasoline refining. But gasoline refining still works without rare earth catalysts, just slightly less efficiently; in fact, at the peak of the 2011 rare-earths price bubble (well after the embargo crisis), some refiners stopped using the rare earth catalysts to save input costs. . . . Companies such as Hitachi Metals that make rare earth magnets (now including in North Carolina) found ways to make equivalent magnets using smaller amounts of rare earths in the alloys. Some users remembered that they did not need the high performance of specialized rare earth magnets; they were merely using them because, at least until the 2010 episode, they were relatively inexpensive and convenient.” Overall, Gholz writes: “Future crises are unlikely to seem so perfectly orchestrated to make the United States and its allies vulnerable: the materials in question may be more prosaic or the country where supplies are concentrated may loom less ominously than China. But even in the apparently most-dangerous case of rare earth elements, the problem rapidly faded—and not primarily due to government action.” Council on Foreign Relations, October 2014. At http://i.cfr.org/content/publications/attachments/Energy%20Report_Gholz.pdf.

Roland Fryer discusses his journey to becoming a K-12 school reformer in “21st Century Inequality: The Declining Significance of Discrimination.” “When do U.S. black students start falling behind? It turns out that development psychologists can begin assessing cognitive capacity of children when they are only nine months old with the Bayley Scale of Infant Development. We examined data that had been collected on a representative sample of 11,000 children and could find no difference in performance of racial groups. But by age two, one can detect a gap opening, which becomes larger with each passing year. By age five, black children trail their white peers by 8 months in cognitive performance, and by eighth grade the gap has widened to twelve months.” Fryer’s focus has been to study successful charter schools, and then to try to apply their lessons. In a group of 20 Houston public schools, including four high schools, with 16,000 students, they found: “When we began, the black/white achievement gap in the elementary schools was about 0.4 standard deviations, which is equivalent to about 5 months. Over the three years, our elementary schools essentially eliminated the gap in math and made some progress in reading. In secondary schools, math scores rose at a rate that would close the gap in roughly four to five years, but there was no improvement in reading. One other significant result was that 100% of the high school graduates were accepted to a two- or four-year college.” The essay appears in *Issues in Science and Technology*, Fall 2014, at http://issues.org/31-1/21st-century-inequality-the-declining-significance-of-discrimination. The article is an edited version of the Henry and Bryna David Lecture, which Fryer delivered at the National Academy of Sciences on April 29, 2014, and can be watched at http://sites.nationalacademies.org/DBASSE/DBASSE_088044.

CONSIDERATIONS FOR THOSE PROPOSING TOPICS AND PAPERS FOR JEP

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.

Philosophy and Style

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of *JEP* articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author.** The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

JEP does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-U.S. country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives*. Our stock in trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship. If you are not familiar with this journal, it is freely available on-line at <http://e-*JEP*.org>.

Guidelines for Preparing JEP Proposals

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP*. However, given the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and authors an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

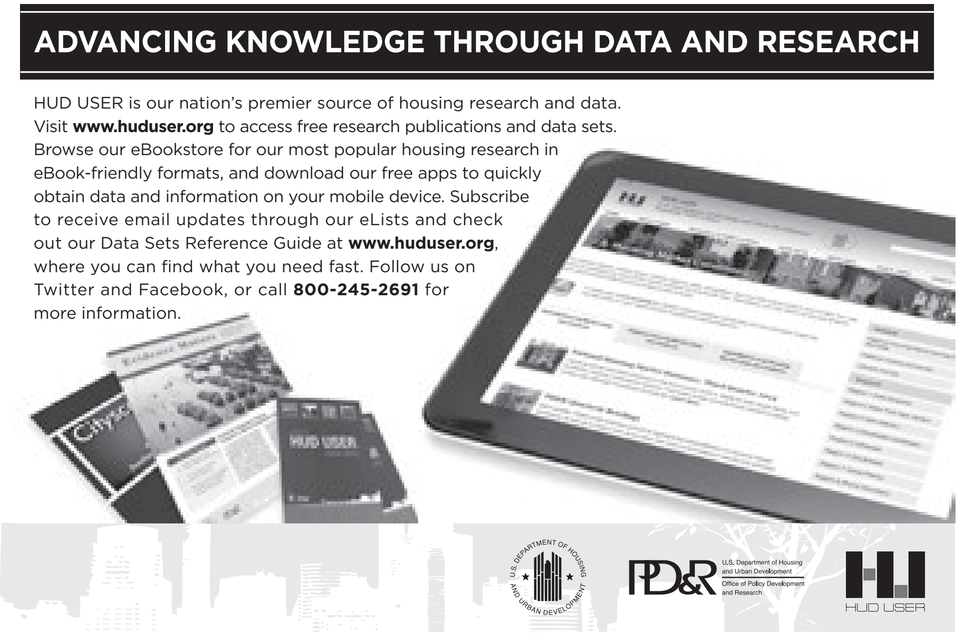
Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future articles for *JEP* may be sent (electronically please) to the assistant editor, Ann Norman, at <anorman@JEPjournal.org>. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

Guidelines for Empirical Papers Submitted to JEP

The *JEP* is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1. The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.
2. In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.
3. The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.



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*

Spring 2015, Volume 29, Number 2

Symposia

The Bailouts of 2007–2009

Disability Insurance

Articles

AMERICAN
ECONOMIC
ASSOCIATION