CATO PAPERS ON PUBLIC POLICY

Volume 3, 2013–2014

CATO PAPERS ON PUBLIC POLICY

Volume 3, 2013–2014

CATO INSTITUTE

Washington, D.C.

Cato Papers on Public Policy, ISBN: 978-1-939709-10-3 (Paperback); ISBN: 978-1-939709-11-0 (Digital). *CPPP* is published by the Cato Institute, a nonprofit, nonpartisan 501(c)(3) research organization based in Washington, D.C.

Correspondence regarding subscriptions, changes of address, procurement of back issues, advertising and marketing matters, and so forth should be addressed to:

Publications Department
Cato Institute
1000 Massachusetts Avenue, N.W.
Washington, D.C. 20001

All other correspondence, including requests to quote or reproduce material, should be addressed to the managing editor.

Printed in the United States of America.

Cato Institute
1000 Massachusetts Avenue, N.W.
Washington, D.C. 20001
www.cato.org

Publications Director: *David Lampo*
Marketing Director: *Robert Garber*
Circulation Manager: *Alan Peterson*
Cover: *Jon Meyers*

Contents

ARTICLES

Off-Balance-Sheet Federal Liabilities

James D. Hamilton

ABSTRACT

Much attention has been given to the recent growth of the U.S. federal debt. This paper examines the growth of federal liabilities that are not included in the official debt numbers. Those numbers take the form of implicit or explicit government guarantees and commitments. The five major categories surveyed are support for housing, other loan guarantees, deposit insurance, Federal Reserve actions, and government trust funds. The total dollar value of notional off-balance-sheet commitments came to $70 trillion as of 2012, or six times the size of the reported on-balance-sheet debt. This paper reviews the potential costs and benefits of those off-balance-sheet commitments and their role in precipitating or mitigating the financial crisis of 2008.

James D. Hamilton is professor of economics at the University of California, San Diego.

The author thanks Marjorie Flavin, Robert Hall, Douglas Holtz-Eakin, Jeffrey Miron, Phillip Swagel, and seminar and conference participants at the University of California, San Diego, and the Cato Institute for helpful comments on this paper.

Off-Balance-Sheet Federal Liabilities

1. INTRODUCTION

U.S. federal debt has exploded in recent years, growing from $5 trillion (or 36 percent of gross domestic product [GDP]) in 2007 to an estimated $12 trillion (72 percent of GDP) by the end of 2013. Making the interest payments on that debt poses a permanent burden on future taxpayers. At the moment, interest rates are at historic lows, with the government paying an average rate of only 2 percent to service its outstanding debt.[1] That rate contrasts with an average rate of 4.5 percent paid over 2000–2009 and 6.6 percent over 1990–1999.

Most projections call for interest rates to rise back to more usual historical levels over the next several years. For example, the consensus Blue Chip Financial Forecast anticipates a yield on 10-year Treasury bonds of 4.7 percent by 2017 (Bernanke 2013). Returning to those levels of interest rates or the even higher rates seen on average during the 1990s would mean a doubling or tripling of the government's current annual interest expense, bringing it to 2.8–4.2 percent of GDP, even without further increases in federal debt from now on. The Congressional Budget Office (CBO) currently anticipates that net interest expense will exceed the entire defense budget by 2021.[2]

But the officially reported debt is only one aspect in which current policy has left a burden for future taxpayers. In addition, the government has made a number of implicit and explicit commitments that are not included in the net debt figures just reported, but that could potentially require much larger adjustments in future spending or taxes than those associated with paying interest on the official debt itself. The biggest items in that category come from Social Security

[1] Federal net interest expense in 2012 was $220 billion, which was 2 percent of the $11,281 billion in debt held by the public. Data are from the Congressional Budget Office, "Historical Budget Data," May 2013, http://www.cbo.gov/publication/44197.

[2] Congressional Budget Office, "Updated Budget Projections: Fiscal Years 2013–2023," May 2013, http://www.cbo.gov/sites/default/files/cbofiles/attachments/44172-Baseline2.pdf.

and Medicare, which, if current policy is maintained, will require enormous sacrifices from future taxpayers. Gokhale (2013) and the 2013 Trustees Reports for Social Security and Medicare have provided updates on the size of those commitments. However, a number of other government guarantees and implicit commitments are also quite significant.

This paper tabulates the growth of the major categories of off-balance-sheet commitments, with a particular focus on their possible role in precipitating or mitigating the financial crisis of 2008. Adding all the off-balance-sheet liabilities together, I calculate that total federal off-balance-sheet commitments came to $70.1 trillion as of 2012, or about six times the size of the on-balance-sheet debt. In other words, the budget impact associated with an aging population and other challenges could turn out to have much more significant fiscal consequences than even the mountain of on-balance-sheet debt already accumulated.

This paper begins with an examination of the motivations for off-balance-sheet commitments. Those motivations include prevention of or response to financial crises, subsidization of socially desired activities, and commitments to retirees. I then review five of the most important current sources of federal off-balance-sheet liabilities: support for housing, other loan guarantees, deposit insurance, Federal Reserve actions, and government trust funds. I explore how those commitments have changed over the last six years and their possible role in contributing to or mitigating the financial strains associated with the Great Recession. A final section briefly draws on some other historical episodes to offer some concluding thoughts on how concerned we should be about the growth of off-balance-sheet federal liabilities.

2. REASONS FOR OFF-BALANCE-SHEET COMMITMENTS

There are three traditional motivations for the government to make commitments in addition to the explicit liabilities embodied in outstanding sovereign debt. In this section, I briefly review those motivations.

2.1 Prevention of, and Response to, Financial Crises

Financial institutions like banks function as intermediaries between potential lenders and borrowers. The lenders value such intermediation because they desire to hold their assets in highly liquid form,

such as bank deposits that could be withdrawn at any time. The borrowers have longer-term investment projects with incompletely observed fundamentals. As a result, the bank's loans are inherently illiquid. The financial sector issues short-term liquid liabilities to the lenders and holds long-term, illiquid assets (claims against the borrowers), thereby earning a profit on the interest rate spread.

Diamond and Dybvig (1983) suggest that such arrangements are inherently unstable. If all the bank's depositors were ever to want their money back at the same time, an immediate sale of the bank's assets would involve a loss as a necessary result of their illiquid character. In such an event, depositors would not get all their money back, and a bank run could be self-fulfilling. Diamond and Dybvig show that this problem can be prevented if the government guarantees all deposits on the basis of its taxing authority, which ideally would never need to be invoked in equilibrium.

Other economists have argued that deposit insurance could itself introduce a moral hazard problem, causing banks to make riskier loans than they would have in the absence of a government guarantee; see, for example, Kareken and Wallace (1978) and the surveys in Bhattacharya, Boot, and Thakor (1998) and Santos (2001). Moreover, bank runs could alternatively be prevented by requiring the bank to carry a level of equity—initial capital provided by the bank's owners rather than obtained from depositors or short-term lenders—that is sufficient to cover the losses should the bank be forced to liquidate. This situation involves a potential efficiency loss relative to government deposit insurance in that it requires capital to be tied up as reserves as a contingency against liquidity needs; for discussion, see Diamond and Rajan (2000) and Cooper and Ross (2002).

An alternative role for the government could arise *ex post* in deciding how to respond to a financial panic if one were to occur. If banks are forced to sell off their illiquid assets at the height of a financial panic, such "fire sales" would depress the market price of illiquid assets even further. There would be social benefits in such a setting from having the government or central bank offer emergency loans, which the banks would be able to repay over time through a more orderly sale of bank assets. A classic discussion of the role for government lending in such a situation was provided by Bagehot (1873); see Flannery (1996) and Brunnermeier and Sannikov (2012) for modern assessments.

2.2 Subsidization of Socially Desired Activities

It is also possible to make a case for government loans or loan guarantees as a steady-state policy even if bank runs are not a concern. The government is able to borrow at lower rates than any private borrower. To the extent that the government can do this because of private-lending frictions (such as principal-agent problems or credit rationing), it may be welfare improving to have the government directly provide loans or guarantees to many private borrowers. Nevertheless, moral hazard and other problems could again cause such government assistance to be counterproductive; see for example Chaney and Thakor (1985), Smith and Stutzer (1989), and Bencivenga and Smith (1993).

Alternatively, a social planner might choose to direct more capital to certain activities than would be the outcome in an unregulated market if those activities are associated with positive economic externalities. For example, homeownership may be perceived to lower crime rates and increase community involvement, and education could provide a range of external social benefits. Such arguments could serve as possible justification for the large government participation in housing and student loans discussed in Section 3. In that discussion, I will also review the potential downside of such programs.

2.3 Commitments to Retirees

Finally, the government would need to offer retirement benefits to many of its employees in order to compete with private-sector employers for workers. The government's existing commitment to its future retired workers may take the form of off-balance-sheet liabilities. More broadly, the U.S. government has also undertaken limited provision for almost all retired and many disabled workers in the form of the Social Security Trust Fund. Such government involvement in retirement assistance to private-sector workers could be motivated on the basis of a perceived public good. As a nation, we do not want to see the elderly suffer, even if it is a result of their own decision not to save when they were working. Assisting such individuals is a public good in the sense that providing the good for me (giving me the satisfaction that Person X is adequately cared for) does not reduce the benefit to you of that same good. Such programs could be regarded as an off-balance-sheet liability if existing policies commit the government to offering a certain level of assistance to retirees in years to come.

3. KEY U.S. OFF-BALANCE-SHEET COMMITMENTS

In Section 2, I discussed some general reasons why the United States has assumed its many off-balance-sheet obligations. In this section, I examine those obligations specifically and attempt to gauge their size.

3.1 Housing

One of the most important areas of federal off-balance-sheet commitments involves assistance to the housing market. Those programs began in 1934 when Congress established the Federal Housing Administration (FHA), which insures approved mortgages. Such insurance creates a contingent liability of the federal government that is not counted as part of the federal debt and, if the loan is repaid, never shows up on the Treasury's income or balance sheet. Since its inception, the FHA has insured 40 million loans. During 2012, the FHA insured $213 billion in new mortgages, bringing its total portfolio of insured mortgages to $1.3 trillion (U.S. Department of Housing and Urban Development 2012).

In 1938, Congress created the Federal National Mortgage Association, commonly known as Fannie Mae, as a separate entity to purchase the loans that were guaranteed by the FHA. Although originally created through an act of Congress, Fannie Mae has for much of its history had some of the characteristics of a private corporation, with the Federal National Mortgage Association Charter Act of 1954 giving the federal government and private stockholders mixed ownership of the enterprise. In 1968, Fannie Mae was split into two separate entities: the Government National Mortgage Association (Ginnie Mae) was to be an entirely government-owned corporation intended particularly to assist lower-income households, while the remainder (which continued to be referred to as Fannie Mae) was intended to function mostly as a private enterprise separate from the government. In 1970, Congress chartered the Federal Home Loan Mortgage Corporation (Freddie Mac) to serve a similar role and act as a competitor to Fannie Mae.

Because both Fannie Mae and Freddie Mac were originally created by an act of Congress, they are referred to as "government-sponsored enterprises" (GSEs). For a number of reasons, it is difficult to regard them as ever being truly private companies. They were exempt from all state and local taxes other than property taxes, and bankruptcy procedures were never clear. The Federal National Mortgage

Association Charter Act did specify that, in issuing any debt, Fannie Mae "shall insert appropriate language in all of its obligations issued under this subsection clearly indicating that such obligations, together with the interest thereon, are not guaranteed by the United States and do not constitute a debt or obligation of the United States or of any agency or instrumentality thereof other than the corporation" (Section 304.2.b). But lenders to the GSEs did not act as if they actually believed that. For example, a 2001 CBO assessment estimated that the GSEs would have had to pay an average interest rate that would be 41 basis points higher than the rate they actually paid to borrow if lenders did not perceive a significant probability that the U.S. Treasury would step in, if necessary, to prevent a GSE default on the debt. See Haffner (2008) for further discussion.

Fannie and Freddie used the funds borrowed at advantageous rates to purchase significant volumes of new mortgages. As seen in Figure 1, their combined holdings increased by almost $1.3 trillion between 1994 and 2004, or over a 300 percent increase; U.S. nominal GDP increased only 67 percent over that same period. As of the end of 2009, Fannie owed $780 billion in short-term and long-term debt, which it used to finance $736 billion in mortgages that it held at the time.[3] Freddie owed $781 billion and held $718 billion in mortgages.[4]

Other GSEs were also using funds borrowed at favorable rates to purchase mortgages. For example, at the end of 2009, the Federal Home Loan Banks held $616 billion in financial assets, and the Farm Credit System held $80 billion.[5] Those four enterprises, along with the Financing Corporation and the Resolution Funding Corporation, had total debt outstanding of $2.7 trillion as of the end of 2009 (Federal Reserve Board, *Flow of Funds*, Table L.1).

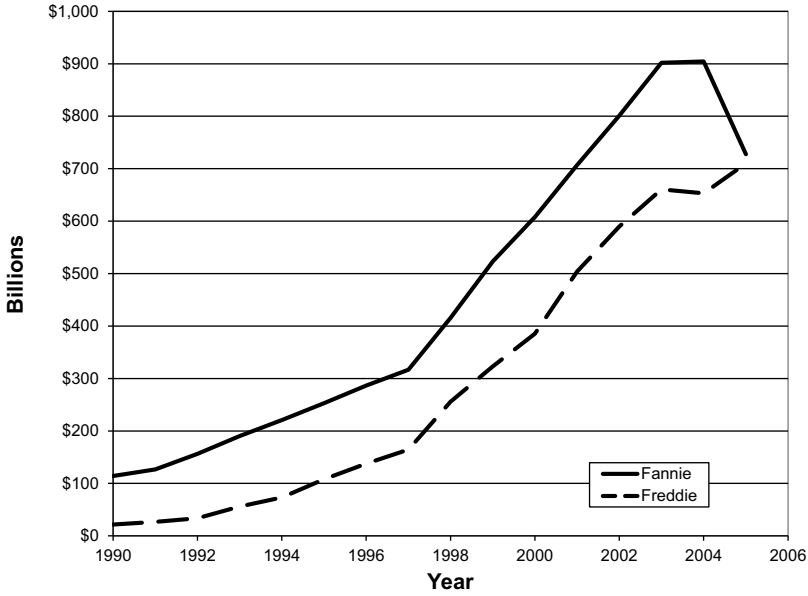
Even more important in dollar terms were the guarantees that Fannie and Freddie provided for a vastly larger volume of loans. The system worked as follows: A private entity would lend its own funds to a household for the purpose of buying a house and would then sell the mortgage to Fannie or Freddie. The GSE would then bundle a set

[3] This number refers to the sum of mortgages and mortgage-backed securities held for trading, available for sale, and held for investment. Data are from 2010 Fannie Mae Annual Report, p. F-3.

[4] Data are from 2010 Freddie Mac Annual Report, p. 175.

[5] See Federal Reserve Board, *Flow of Funds*, Table L.122, "Government-Sponsored Enterprises."

Figure 1
Retained Mortgage Portfolios of Fannie Mae and Freddie Mac
(first quarter, 1990 to third quarter, 2006)



Source: Hamilton (2007b).

of loans into an asset known as a mortgage-backed security (MBS) and sell the bundle to individual investors. Fannie described its business model as follows:[6]

> We support market liquidity by securitizing mortgage loans, which means we place loans in a trust and Fannie Mae MBS backed by the mortgage loans are then issued. We guarantee to the MBS trust that we will supplement amounts received by the MBS trust as required to permit timely payment of principal and interest on the trust certificates. In return for this guaranty, we receive guaranty fees.

Although those guarantees represented a liability of the GSE, they were not counted as part of the GSE's own balance sheet. They were,

[6] From Fannie Mae 2011 Annual Report, p. 22.

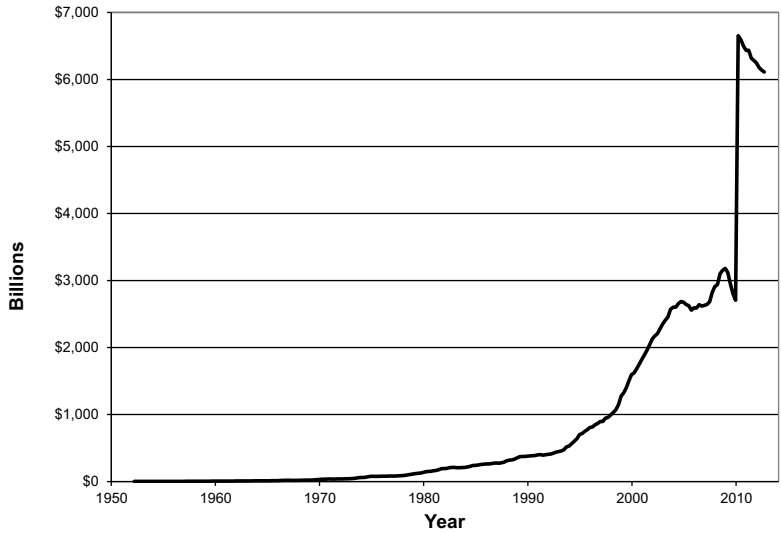
in effect, an off-balance-sheet liability of an enterprise that could itself be regarded as an off-balance-sheet liability of the federal government. As of the end of 2009, the assets in agency-backed and GSE-backed mortgage pools came to $5.4 trillion (Federal Reserve Board, *Flow of Funds*, Table L.1).

The combined net equity of Fannie and Freddie that could be used to honor this guarantee came to only $70 billion as of 2006. That such enterprises would be capable by themselves of actually guaranteeing such a vast sum stretches credulity. As housing prices began to fall after 2006, the GSEs' net equity turned negative, and the federal government placed both Fannie and Freddie into conservatorship. Beginning in 2010, the guarantees began to be reported as on-balance-sheet liabilities of the GSEs (see Figure 2).

Having been originally created through an act of Congress, and with the federal government today being the sole owner of Fannie and Freddie, it seems appropriate to consider both the direct debt obligations of the GSEs and other government housing agencies, as well as their outstanding mortgage guarantees, as an off-balance-sheet liability of the federal government. It should be recognized that such liabilities do not have the same status as the direct debt obligations of the Treasury itself. For one thing, there are some offsetting assets, namely, the mortgages held outright. The value of the mortgages would never fall to zero, so using the notional exposure is a significant overstatement of the conceivable net outlays that would be required from the federal government to honor these commitments. Nevertheless, it seems a useful exercise to calculate the total notional value of these off-balance-sheet debts and guarantees.

Line 2 of Table 1 reports the par value of the outstanding value of the direct debt of Fannie, Freddie, Federal Home Loan Banks, Farm Credit System, Federal Agricultural Mortgage Corporation (Farmer Mac), Financing Corporation, and Resolution Funding Corporation at the end of various fiscal years, while line 3 reports the notional value of the separate mortgage guarantees issued by Fannie, Freddie, Ginnie, and the Farmers Home Administration, plus mortgages held outright by the Federal Financing Bank. The big jump in line 2 and the fall in line 3 after September 2008 represent the accounting decision to move Fannie's and Freddie's guarantees on balance sheet. Lines 2 and 3 sum to $7.5 trillion as of the end of fiscal year 2012, two-thirds as big as the entire stock of Treasury debt held by the

Figure 2
Total On-Balance-Sheet Liabilities of GSEs
(first quarter, 1990 to third quarter, 2006)



Source: Federal Reserve Board, *Flow of Funds*, Table L.122.
Note: Includes Federal Home Loan Banks, Fannie Mae, Freddie Mac, Federal Agricultural Mortgage Corporation (Farmer Mac), Farm Credit System, the Financing Corporation, and Resolution Funding Corporation. The Student Loan Marketing Association (Sallie Mae) was included until it was fully privatized in the fourth quarter of 2004. Beginning in the first quarter of 2010, almost all Fannie and Freddie mortgage pools are consolidated on Fannie's and Freddie's balance sheets.

public. Note that we have left out of this calculation the $1.3 trillion in loan guarantees[7] issued by the Federal Housing Administration, Veterans Housing Benefit Programs, and Rural Housing Service, on the assumption that most of those loans ended up being held as assets or part of the guaranteed pools of the GSEs and agencies that are already included in lines 2 and 3 of Table 1.

Panel A of Figure 3 provides historical perspective by plotting the combined mortgage debt either held outright by the GSEs or held in

[7] This number comes from Government Accountability Office, "Financial Statements of the United States Government: Notes to Financial Statements," Fiscal Year 2012.

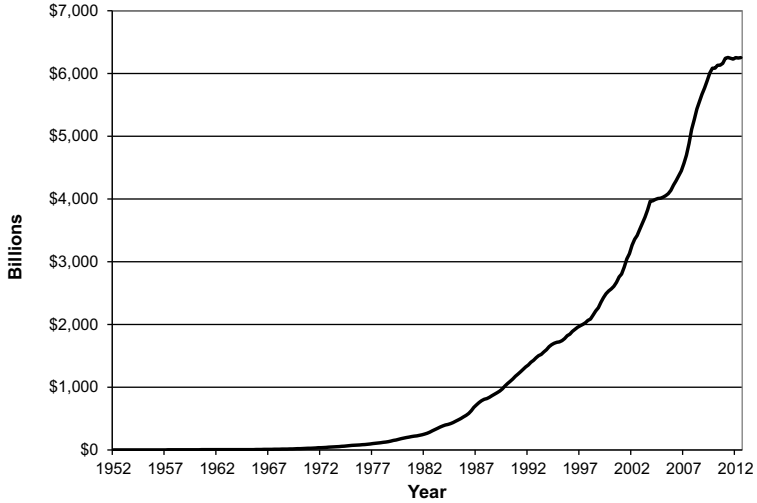
Table 1
Treasury Debt Held by the Public and Contribution of Guarantees and Obligations of GSEs to Off-Balance-Sheet Federal Liabilities (fiscal year, in billions of dollars)

	2006	2008	2010	2012
1. Treasury debt held by public	$4,867	$5,837	$9,052	$11,299
2. Direct debt obligations of government-sponsored enterprises	$2,618	$3,153	$6,498	$6,112
3. MBS guaranteed by agencies and GSEs	$3,769	$4,883	$1,096	$1,408
4. Total off-balance-sheet commitment to housing	**$6,386**	**$8,036**	**$7,594**	**$7,520**
5. Student loan guarantees	$321	$405	$382	$285
6. Less: student loan guarantees already imputed to on-budget federal debt	$0	−$11	−$137	−$165
7. Other loan guarantees	$147	$153	$174	$205
8. Combined GSE debt and loan guarantees	**$6,854**	**$8,583**	**$8,013**	**$7,845**

Sources and Notes: **Line 1:** Public debt securities and agency securities held by the public from Table FD-1 ("Summary of Federal Debt"), *Treasury Bulletin*, December 2008 and December 2012. **Line 2:** Short-term and long-term debt issued by Freddie, Fannie, Federal Home Loan Banks, Farm Credit System, Farmer Mac, Financing Corporation, and Resolution Funding Corporation. After September 2008, this debt also includes MBS guarantees of Fannie and Freddie. From line 14 ("credit market debt owed by GSEs"), Table L.1 ("Credit Market Debt Outstanding"), *Flow of Funds*, Federal Reserve Board. **Line 3:** Mortgages held in pools from Ginnie, Freddie, Fannie, and Farmers Home Administration plus mortgages held by Federal Financing Bank. After 2008, Fannie and Freddie MBS are included in line 2 but not line 3. From line 28 ("agency and GSE-backed mortgage pools"), Table L.217 ("Total Mortgages"), *Flow of Funds*, Federal Reserve Board. **Line 4:** Sum of lines 2 and 3. **Line 5:** Principal amount guaranteed by the government for Federal Family Education Loans, from Government Accountability Office, "Financial Statements of the United States Government: Notes to Financial Statements," indicated fiscal years. **Line 6:** Debt owed by Department of Education to Department of the Treasury associated with Federal Family Education Loans. From columns (3)–(6) of Table 2. **Line 7:** Principal amounts guaranteed by the government for programs other than FHA, Rural Housing Services, Veterans Housing Benefit Programs, and education loans, from Government Accountability Office, "Financial Statements of the United States Government: Notes to Financial Statements," indicated fiscal years. **Line 8:** Sum of lines 4–7.

Figure 3
Mortgage Debt Held by GSEs or in Agency-Backed or GSE-Backed Mortgage Pools
(first quarter, 1952 to third quarter, 2012)

Panel A. Amount held



Panel B. Amount as a percentage of U.S. GDP

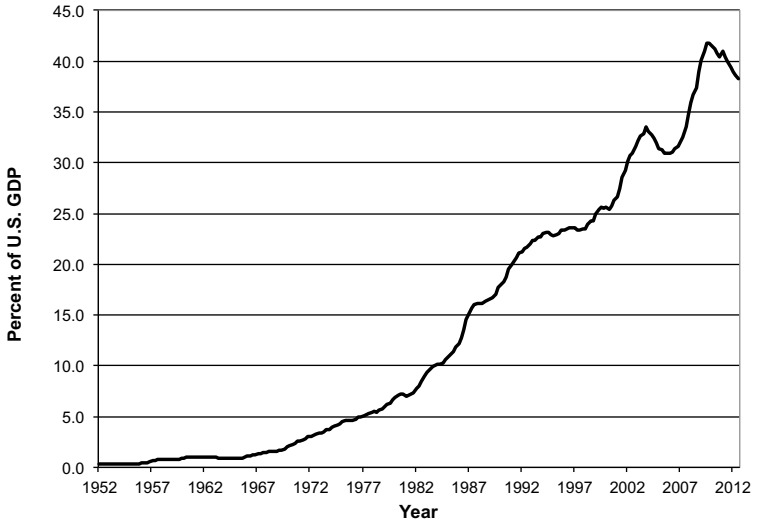
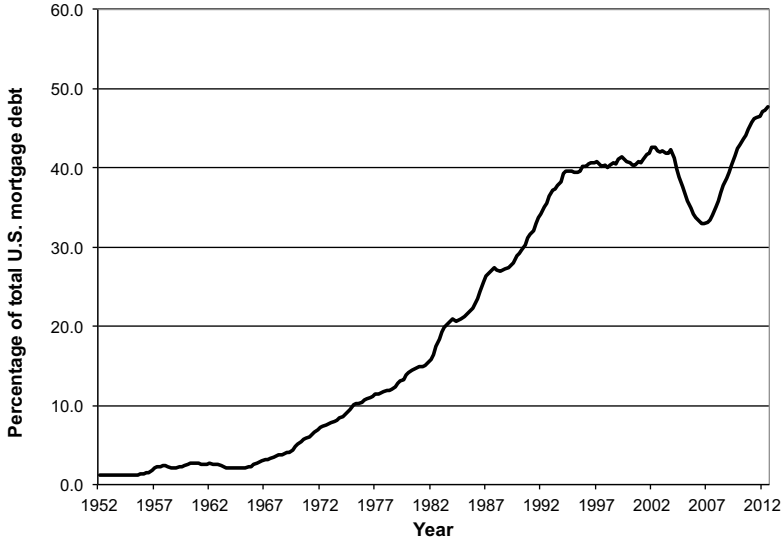


Figure 3
(*continued*)

Panel C. Amount as a percentage of all U.S. mortgages



Source: Federal Reserve Board, *Flow of Funds*, Table L.217.

mortgage pools that were guaranteed by the GSEs or other federal agencies. That combined debt nearly quadrupled between 1991 and 2006, and it more than doubled as a percentage of GDP over those 15 years (see Panel B of Figure 3).

What have been the public benefits of that off-balance-sheet commitment? One goal of those policies was to increase the homeownership rate. There was indeed a significant rise in homeownership from 64 percent in 1994 to 69 percent by 2005. However, those gains proved to be temporary, as they were mostly wiped out by the housing crash and Great Recession (see Figure 4).

Before the crash, however, the U.S. housing boom was pretty spectacular. Household mortgage debt grew significantly faster than GDP in the decade leading up to the crash (Figure 5). U.S. real house prices, which according to data by Shiller (2000) had been relatively

Figure 4
U.S. Homeownership Rate
(1965–2012, fourth quarter of each year)

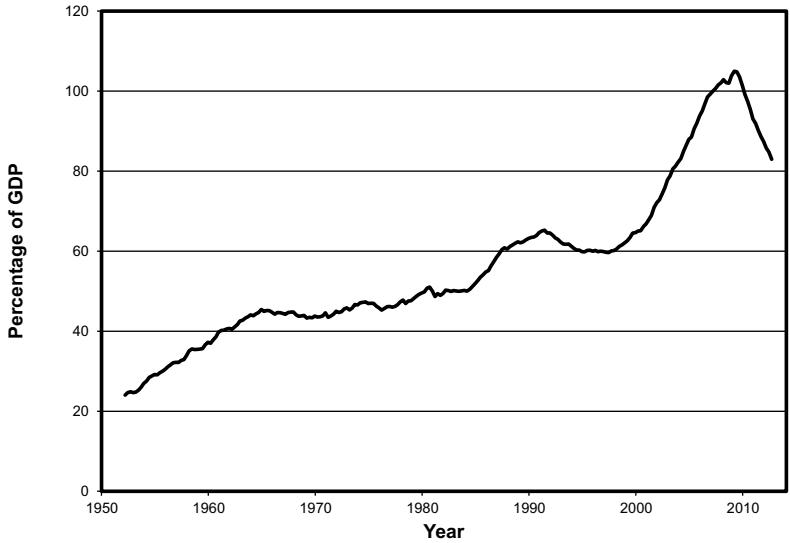
Source: United States Census Bureau, Historical Table 14, http://www.census.gov/housing/hvs/data/histtabs.html.

stable for a century, nearly doubled within the span of a decade before crashing just as dramatically (Figure 6).

That the tremendous growth in implicit federal guarantees contributed to the housing bubble seems difficult to deny. However, it is important to recognize that an even bigger contribution came from outside the GSEs and federal agencies. As seen in Panel C of Figure 3, despite the rapid growth of mortgage debt held or guaranteed by the GSEs, the fraction of mortgage debt associated with the GSEs climbed only modestly between 1995 and 2003, after which it began to decline as a result of an even faster growth in private-label MBS.[8] Those private-label MBS represented an adaptation of the Fannie and Freddie model. As with the GSEs, a mortgage originator

[8] See Ashcraft and Schuermann (2007) for a description of private-label MBS.

Figure 5
Total Mortgage Debt as a Percentage of GDP
(first quarter, 1952 to third quarter, 2012)

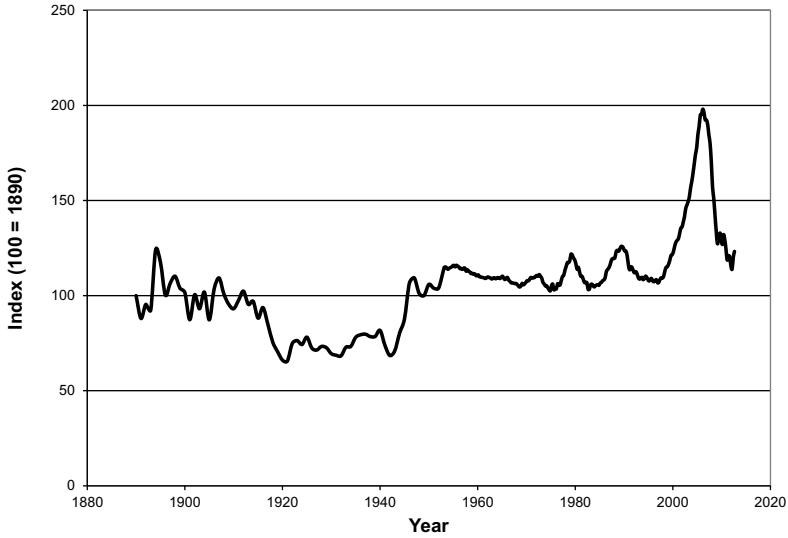


Source: Federal Reserve Board, *Flow of Funds*, Table L.217.

would make the initial loan to a household. But rather than sell the mortgage to Fannie or Freddie, the originator would sell the loan to a private loan aggregator, such as New Century Financial (which went bankrupt in 2007) or Countrywide Financial (which became distressed and was purchased by Bank of America in 2008). The aggregator would collect a group of thousands of mortgages into a pool, and the income flows resulting from interest payments on the pool were then assigned to tranches, with the more senior tranches guaranteed to be paid first.

The theory was that the diversification provided by pooling, along with the protection provided by a more senior position in the tranche system, could make such securities safe even without an explicit guarantee from a GSE. That belief gained acceptance despite the fact that the creditworthiness of the borrowers was substantially worse than that required for securitization by a GSE.

Figure 6
Shiller's Real House Price Index
(1890–2012)

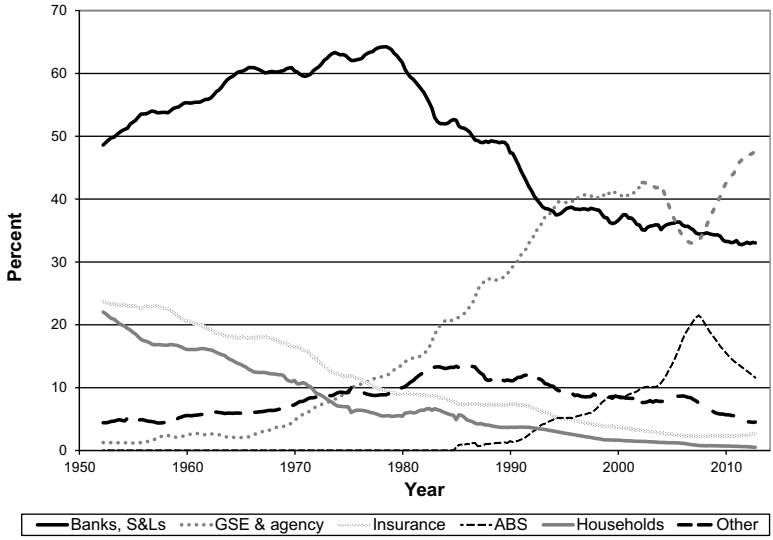


Source: Shiller (2000), as updated at http://www.irrationalexuberance.com/index.htm.

Figure 7 summarizes changing funding sources for U.S. mortgages over time. During the 1980s, GSE pools gradually displaced private banks and savings and loans as the dominant provider of mortgage funds in the United States. During the 1990s, the GSE share increased only modestly as private-label MBS grew rapidly. The private-label MBS were far more important than the GSEs in the explosion of U.S. mortgage debt in the 2000s.

One could, nevertheless, argue that the implicit guarantees associated with the GSE and agency mortgages made an indirect contribution to the explosion of private-label MBS. The primary risk for both private and GSE MBS was that of a crash in real-estate prices. As long as prices kept rising, even the poor-quality subprime loans had very low default rates because borrowers could refinance at a profit, thus turning their initial position of zero or even negative net equity into a handsome profit. As noted earlier, the potential federal

Figure 7
Percentage of U.S. Mortgage Debt Financed by Different Sectors
(first quarter, 1952 to third quarter, 2012)



Source: Federal Reserve Board, *Flow of Funds*, Table L.217.
Note: *ABS:* Mortgages held in the form of privately issued asset-backed securities. *Banks, S&Ls:* Sum of U.S.-chartered depository institutions, foreign banking offices in the United States, in U.S.-affiliated areas, and unions. *GSE & agency:* Sum of mortgages held by GSEs and in agency-backed and GSE-backed mortgage pools. *Households:* Mortgages owned by U.S. household sector. *Insurance:* Sum of mortgages held by property-casualty insurance companies and life insurance companies. *Other:* All other mortgage holdings.

losses from a major real-estate meltdown were quite significant. Some might have believed that, insofar as the government would do whatever necessary to prevent significant losses on the GSE debt and guarantees, the same policies would help the issuers of private MBS keep their noses above water (Hamilton 2007a).

3.2 Other Federal Loan Guarantees

There are a number of important federal loan guarantee programs in addition to those involving housing. One of the most important

programs involves loans for post-high-school education. Lee (2013) estimates from household-level credit data that the outstanding stock of student loans grew from a little over $300 billion in 2004 to nearly $1 trillion by 2012. Federal loans and loan guarantees have played a key role in funding that explosion of student debt.

The U.S. Department of Education has historically provided assistance through two programs. Under the Federal Family Education Loan Program (FFEL), the Department of Education guaranteed certain loans issued by state governments or private entities. There were $493 billion in outstanding student loan guarantees at the program's peak in 2009.[9] The program was discontinued in July 2010 with the intention of replacing it with greater reliance on direct loans from the Department of Education.

The financing for the latter works as follows: The U.S. Treasury borrows from the public through its general auctions and earmarks some of those funds for the Department of Education. One of the things the Department of Education then does with those funds is make direct loans to students. Treasury borrowings earmarked for the Department of Education grew from $104 billion at the end of FY 2007 to $714 billion at the end of FY 2012.[10] That $714 billion is already included in the $11,299 billion Treasury debt reported in line 1 of Table 1 as being held by the public at the end of FY 2012, and indeed it accounts by itself for 6 percent of the total publicly held debt. It is thus already on balance sheet as far as the reported debt is concerned. Interestingly, this activity is off balance sheet from the perspective of the reported deficit. Treasury borrowing earmarked for student loans is one of the reasons that the reported growth in publicly held debt in recent years has been bigger than the reported deficit.

Insofar as that borrowing for student loans has been associated with the acquisition of an asset (namely, the promise by the students to repay the loans), one can make a case that the $714 billion in Department of Education debt should also be moved off balance

[9] Government Accountability Office, "Federal Student Loan Programs: Opportunities Exist to Improve Audit Requirements and Oversight Procedures," Report no. GAO-10-668, July 2010.

[10] This sum is reported in Table FD-7 of the *Treasury Bulletin* (U.S. Treasury) as an entry in the category "Treasury holdings of securities issued by government corporations and other agencies."

sheet. On the other hand, there is growing evidence of problems with those loans. Lee's (2013) analysis of micro data suggests that of the nearly $1 trillion in outstanding student loans, 44 percent of loans are not being asked to make any payment yet, and for more than two-thirds of those, the outstanding balance due is growing. Of the 56 percent that are supposed to be in repayment, about one-third of loans are delinquent. Presumably, the Department of Education has much more information on exactly why so little is being collected on outstanding students loans, though I have not found an official report of the relevant data.

However, the financial notes to the Department of Education's annual reports of the Federal Student Aid (FSA) programs contain some interesting details. Of the $714 billion in Treasury borrowing at the end of FY 2012 that had been earmarked for the Department of Education, only $549 billion actually went to direct student loans.[11] Another $165 billion was associated with FFEL, which, as noted above, had been discontinued in July 2010.

It is interesting to trace how that sum earmarked for FFEL has grown and continues to grow over time. In August 2008, the Department of Education began buying certain FFEL loans directly under programs such as the Loan Participation Purchase Program and Loan Purchase Commitment Program "to ensure credit market disruptions did not deny eligible students and parents access to federal student loans for the 2008–09 academic year" (FSA 2012 *Annual Report*, p. 10). The dollar value of loans purchased in those programs is summarized in Table 2. Although they are described in the FSA passage just cited as if they were a temporary response to the financial instability in the fall of 2008, the balances have declined very little in the four years since then, and they still stand at $119 billion at the end of FY 2012. Even more interesting is the category for "Guaranteed Loan Program." New loans extended in that category since October 2010 totaled $32 billion; recall that the guaranteed loan program officially ended in July 2010.

I could find little explanation for those entries provided in the financial statements. The natural inference is that the Department of Education has been using funds borrowed from the Treasury to buy up nonperforming guaranteed student loans not only during

[11] See Note 10: Debt, in FSA 2012 Annual Report.

Table 2
Treasury Debt Owed by the Department of Education (totals and contributions of separate components as of end of the indicated fiscal year, in billions of dollars)

	(1)	(2)	(3)	(4)	(5)	(6)
Fiscal Year	**Treasury Debt Owed by Dept. of Education**	**Direct Loan Program**	**Loan Purchase Commitment**	**Loan Participation Program**	**Guaranteed Loan Program**	**Other**
2005	$104	$104	$0	$0	$0	$0
2006	$105	$105	$0	$0	$0	$0
2007	$104	$104	$0	$0	$0	$0
2008	$128	$117	$0	$11	$0	$0
2009	$235	$154	$25	$54	$1	$1
2010	$374	$237	$45	$80	$11	$1
2011	$546	$392	$44	$79	$29	$2
2012	$714	$549	$42	$77	$43	$3

Sources: Financial notes to *Federal Student Aid Annual Reports*, 2010 and 2012, U.S. Department of Education.

the financial crisis but also every year since. To the extent that is the case, the Department of Education has been gradually moving those obligations from the category of off-balance-sheet liabilities (debt guarantees) to on-balance-sheet liabilities (debt owed by the U.S. Treasury to the public).

Lines 5 and 6 of Table 1 summarize the net implications of the student loan program for total off-balance-sheet federal liabilities. Line 5 is an estimate of the outstanding student loan guarantees. Line 6 subtracts out that portion of guaranteed debt that I calculate has now de facto been included in the officially reported on-balance-sheet debt of the U.S. Treasury.

The Government Accountability Office recognizes a few other categories of explicit loan guarantees as official off-balance-sheet liabilities of the U.S. government. The biggest of them are small business loans and loans from the Export-Import Bank of the United States. Those loans added $205 billion to the off-balance-sheet total, as seen in line 7 of Table 1.

3.3 FDIC

The Federal Deposit Insurance Corporation (FDIC) is a government corporation that was created as part of the Banking Act of 1933. The FDIC's role was to insure small depositors against losses if their banks became insolvent. As noted in Section 2.1, in the Diamond-Dybvig (1983) model such insurance could help prevent bank runs, which had been a significant problem in the United States in the initial years of the Great Depression (1929–1933). Deposits insured by the FDIC grew from $2.8 trillion in 1990 to $7.4 trillion at the end of 2012 (see Panel A of Figure 8).

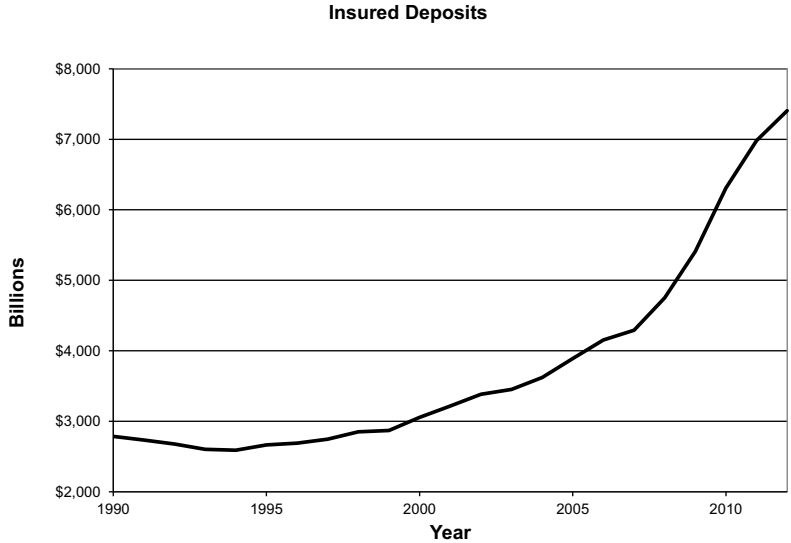
The insurance is funded by a fee on banks. As of the end of 2012, the Deposit Insurance Fund had $33 billion in assets, primarily in the form of debt obligations from the U.S. Treasury that are not included in the $11.3 trillion debt held by the public. Even if the $33 billion represented assets other than the future taxation authority of the Treasury, that equity alone would hardly be sufficient to cover losses should there be a major nationwide bank panic. The Competitive Equality Banking Act of 1987 reaffirmed that "deposits up to the statutorily prescribed amount in federally insured depository institutions are backed by the full faith and credit of the United States." For that reason, those insured deposits are added in our tally of off-balance-sheet federal liabilities in line 1 of Table 3.

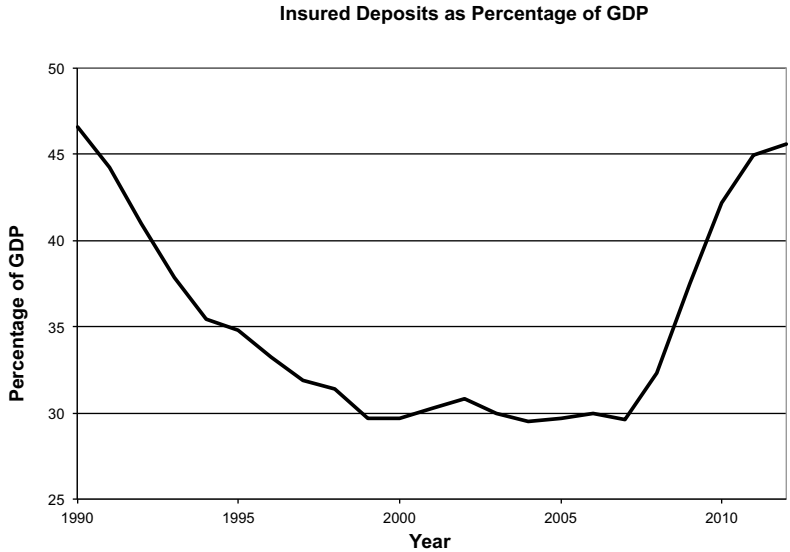
As seen in Panel B of Figure 8, FDIC-insured deposits declined as a percentage of GDP up until 2007, but they climbed swiftly during and after the financial crisis. The Diamond-Dybvig logic suggests that large depositors would have the same incentives to withdraw funds as small depositors, meaning that FDIC insurance of small depositors might not be enough to prevent bank runs. On October 3, 2008, Congress raised the limit on deposit insurance from $100,000 to $250,000, and on October 14, the FDIC instituted the Temporary Liquidity Guarantee Program, which consisted of two components.

One component was the Transaction Account Guarantee Program, which "guaranteed in full all domestic noninterest-bearing transaction deposits, low-interest NOW [negotiable order of withdrawal] accounts, and Interest on Lawyers Trust Accounts . . . held at participating banks and thrifts."[12] The program was

[12] Federal Deposit Insurance Corporation, "Temporary Liquidity Guarantee Program," February 27, 2013, http://www.fdic.gov/regulations/resources/TLGP/index.html.

Figure 8
Total FDIC-Insured Deposits (1990–2012)





Source: FDIC, "Statistics at a Glance," December 2012, http://www.fdic.gov/bank/statistical/stats/index.html.

Table 3
Contribution of FDIC and Federal Reserve to Off-Balance-Sheet Federal Liabilities
(as of end of indicated calendar year, in billions of dollars)

	2006	2008	2010	2012
1. Insured deposits	4,154	4,751	6,308	7,406
2. Plus: debt guarantee program	0	224	267	0
3. Net contribution of FDIC	**4,154**	**4,975**	**6,575**	**7,406**
4. Reserve balances held by member banks	6	856	1,019	1,533
5. Less: Treasury securities held by Federal Reserve	−779	−476	−1,016	−1,657
6. Less: agency debt held by Federal Reserve	0	−20	−147	−77
7. Less: MBS debt held by Federal Reserve	0	0	−992	−927
8. Net contribution of Federal Reserve	**−773**	**360**	**−1,136**	**−1,128**

Sources: **Line 1:** FDIC-insured deposits, from FDIC "Statistics at a Glance," December 2012, http://www.fdic.gov/bank/statistical/stats/index.html. **Line 2:** From FDIC, *Annual Reports*, 2008 and 2010. **Line 3:** Sum of lines 1–2. **Lines 4–7:** From Federal Reserve Release H.4.1 ("Factors Affecting Reserve Balances"), balance as of the last Wednesday of calendar year. **Line 8:** Sum of lines 4–7.

initially intended to expire at the end of 2009. It was later extended twice—to run through the end of 2010—and was replaced by the Dodd-Frank Act with a full guarantee that ran through the end of 2012. Deposits that were insured under Dodd-Frank in excess of the $250,000 limit grew from $854 billion at the end of 2010 to $1,492 billion at the end of 2012.[13] Note that this number is included in the $7,406 billion in total FDIC-insured deposits as of the end of 2012 mentioned earlier. That coverage ended on December 31, 2012, so that the current figure for total FDIC-insured deposits would be approximately $5.9 trillion.

The second component of the Temporary Liquidity Guarantee Program was an FDIC guarantee on certain senior unsecured debt issued between October 14, 2008, and October 31, 2009. The FDIC was guaranteeing an additional $346 billion in debt through

[13] Federal Deposit Insurance Corporation, *Quarterly Banking Profile*, fourth quarter, 2012, Table 1.

the program at its peak. Those guarantees ended on December 31, 2012. The FDIC reports that total fees collected under the Temporary Liquidity Guarantee Program exceeded expenses, and bank runs by depositors were never experienced during the financial crisis.[14]

To summarize, FDIC guarantees in themselves are currently about half as large as the officially reported on-balance-sheet federal debt. However, the stresses of the most recent financial crisis were not enough to cause those guarantees to result in direct cash outflows from the U.S. Treasury, and the program seems to have worked in this instance as intended.

3.4 Federal Reserve

The Federal Reserve System was created by an act of Congress in 1913. Although it is a government entity, it maintains a separate balance sheet of assets and liabilities from the U.S. Treasury. Up until 2007, the Fed's contributions to net government indebtedness were quite straightforward. As the housing decline threatened more financial institutions, the Federal Reserve came to play a much more active role in financial markets. In this subsection, I review those recent developments and discuss their implications for overall federal liabilities.

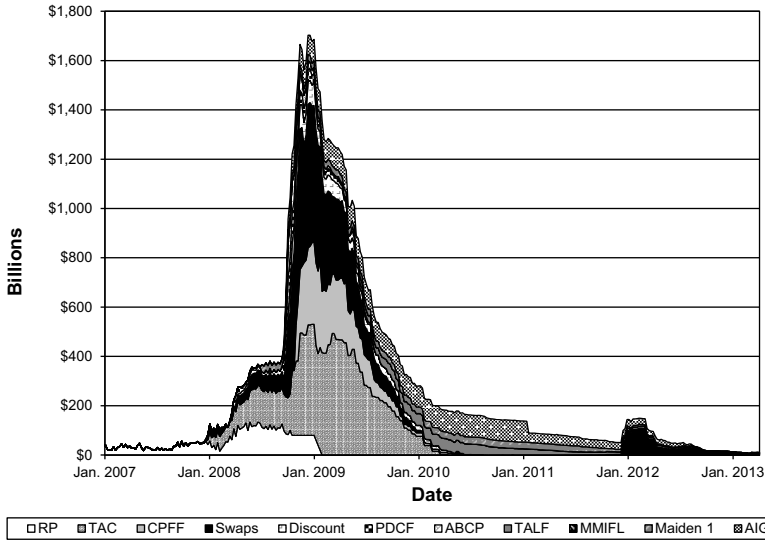
Before 2007, the primary actions of the Federal Reserve consisted of simple open-market operations in which the Fed would buy a debt obligation previously issued by the Treasury, paying for it by crediting the account that the seller maintained with the Federal Reserve System. Those newly created Federal Reserve deposits (which I will refer to as "reserves") represent claims on the Federal Reserve that the bank could, if it wished, subsequently ask to be delivered in the form of currency. Historically, the volume of reserves was very small, averaging $10 billion in 2006. New Fed purchases of Treasury bills were usually followed by conversion of the newly created reserves into currency. One can thus think of the primary function of the Federal Reserve over this period to be swapping one federal liability (T-bills) for another (currency held by the public). The Fed would also occasionally create reserves in order to make short-term loans to banks through the discount window or through a repurchase

[14] Federal Deposit Insurance Corporation, "Temporary Liquidity Guarantee Program," February 27, 2013, http://www.fdic.gov/regulations/resources/TLGP/index.html.

agreement (repo), where the latter can be viewed as a collateralized short-term loan from the Fed. Those loans also were typically quite small, averaging $0.4 billion and $26 billion, respectively, in 2006.

As financial conditions deteriorated in 2008, the Fed made much greater use of existing lending facilities as well as introduced a number of new emergency programs, as seen in Figure 9. Most important

Figure 9
Fed Emergency Lending
(seasonally unadjusted, from January 1, 2007, to April 3, 2013)



Source: Federal Reserve. "Factors Affecting Reserve Balances." *H41 Release*. 2007–2013.
Note: *ABCP:* Loans extended to Asset-Backed Commercial Paper Money Market Mutual Fund Liquidity Facility. *AIG:* Sum of credit extended to American International Group Inc. plus net portfolio holdings of Maiden Lane II and III plus preferred interest in AIA Aurora LLC and ALICO Holdings LLC. *CPFF:* Net portfolio holdings of LLCs funded through the Commercial Paper Funding Facility. *Discount:* Sum of primary credit, secondary credit, and seasonal credit. *Maiden 1:* Net portfolio holdings of Maiden Lane LLC. *MMIFL:* Net portfolio holdings of LLCs funded through the Money Market Investor Funding Facility. *PDCF:* Loans extended to primary dealer and other broker-dealer credit. *RP:* Repurchase agreements. *Swaps:* Central bank liquidity swaps. *TAC:* Term auction credit. *TALF:* Loans extended through Term Asset-Backed Securities Loan Facility plus net portfolio holdings of TALF LLC.

among those programs were currency swaps (agreements in which the Fed temporarily made dollar loans to foreign central banks, collateralized with foreign currency), term auction credit (which supplemented traditional discount borrowing with longer-term collateralized loans to depository institutions), and the Commercial Paper Funding Facility, through which the Fed financed unsecured and asset-backed commercial paper. Some analysts have greatly exaggerated the size of those programs by adding together loans made by the Federal Reserve at different points in time.

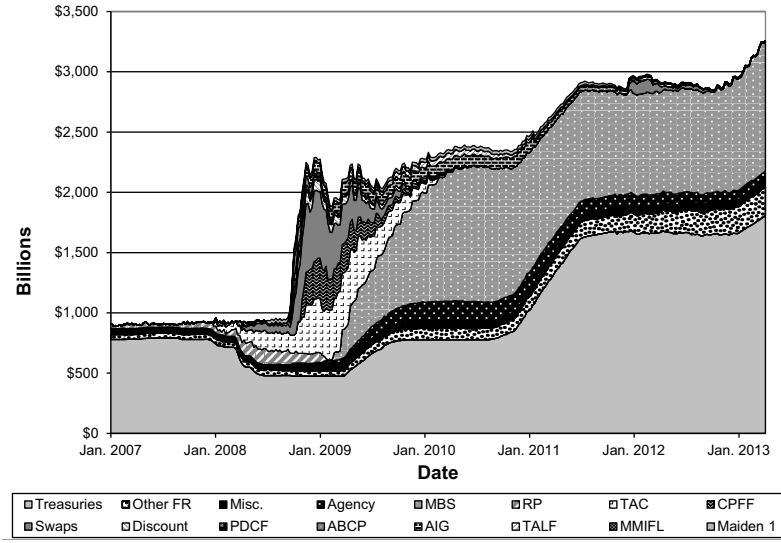
Such accounting is clearly inappropriate, since it would count an overnight loan that is rolled over each day for a month as involving 30 times as much lending as would a single 30-day loan for the exact same amount to the same institution. The accurate measure of the Fed's total exposure through the emergency lending programs is given by the total outstanding loans as of any indicated date, which corresponds to the height of the top line plotted in Figure 9. Total emergency lending reached a maximum of $1,703 billion on December 17, 2008, at which point currency swaps outstanding were $583 billion; term auction credit, $448 billion; commercial paper lending, $319 billion; and all other emergency lending, $353 billion.

As financial conditions eased, the Fed wound down loans and sold off assets associated with all of the programs shown in Figure 9. As a result, the Fed and the Treasury ended up making a profit, with receipts from interest and loan repayments exceeding the total amount lent (Hamilton 2012).

Although those programs were wound down, the Fed's balance sheet did not return to its pre-crisis values, but instead subsequently grew considerably through a series of operations popularly referred to as "quantitative easing" or "large-scale asset purchases." The hope of those programs was that by buying a large enough volume of long-term securities, the Fed might be able to lower the long-term interest rate and thereby stimulate aggregate demand (Hamilton and Wu 2012). Total Federal Reserve assets stood at $3,259 billion on April 3, 2013, of which $1,806 billion consisted of Treasury securities, $1,071 billion mortgage-backed securities with federal agency or GSE guarantees, and $382 billion other assets (see Figure 10).

How do those various Federal Reserve actions affect the net liabilities of the combined Treasury-Fed balance sheet? Consider first an open-market purchase of a T-bill. As noted earlier, that purchase

Figure 10
All Federal Reserve Assets
(seasonally unadjusted, from January 1, 2007, to April 3, 2013)



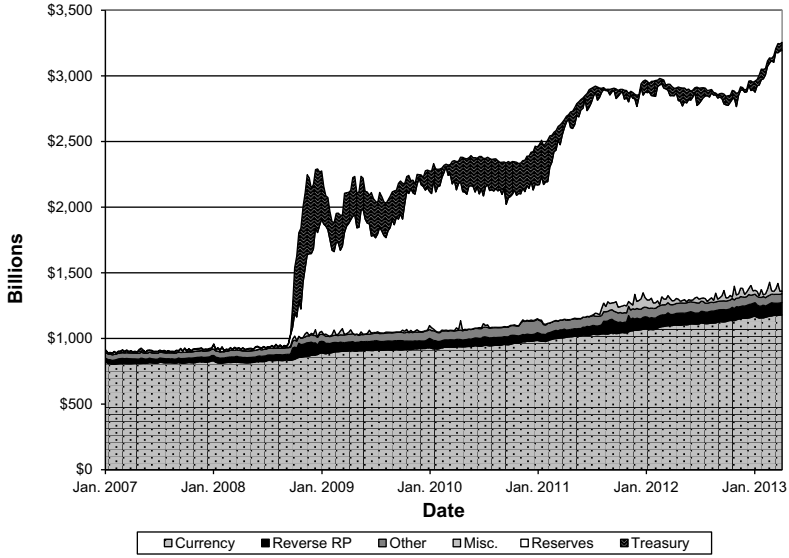
Source: Federal Reserve, "Factors Affecting Reserve Balances," *H41 Release*, 2007–2013.
Note: *ABCP:* Loans extended to Asset-Backed Commercial Paper Money Market Mutual Fund Liquidity Facility. *Agency:* Federal agency debt securities held outright. *AIG:* Sum of credit extended to American International Group Inc. plus net portfolio holdings of Maiden Lane II and III plus preferred interest in AIA Aurora LLC and ALICO Holdings LLC. *CPFF:* Net portfolio holdings of LLCs funded through the Commercial Paper Funding Facility. *Discount:* Sum of primary credit, secondary credit, and seasonal credit. *Maiden 1:* Net portfolio holdings of Maiden Lane LLC. *MBS:* Mortgage-backed securities held outright. *Misc.:* Sum of float, gold stock, special drawing rights certificate account, and Treasury currency outstanding. *MMIFL:* Net portfolio holdings of LLCs funded through the Money Market Investor Funding Facility. *Other FR:* Other Federal Reserve assets. *PDCF:* Loans extended to primary dealer and other broker-dealer credit. *RP:* Repurchase agreements. *Swaps:* Central bank liquidity swaps. *TAC:* Term auction credit. *TALF:* Loans extended through Term Asset-Backed Securities Loan Facility plus net portfolio holdings of TALF LLC. *Treasuries:* U.S. Treasury securities held outright.

swaps one government liability (the T-bill) for another (currency held by the public). Should the latter be viewed as an actual or potential liability of the U.S. government? In the early days of currency issue, it was an actual liability: the public was holding the paper on an explicit understanding that it could be redeemed for gold or silver on demand or at a future announced fiscal date. In the modern era, the public appears willing to hold currency indefinitely, though one could imagine circumstances in which a drop in currency demand might force the Federal Reserve to sell some assets (in effect, redeem the currency for other assets) in order to prevent inflation. For example, Judson (2012) estimates that about half the growth in U.S. currency demand between 1988 and 2011 came from holders outside the United States. Similarly, if the Federal Reserve were to take a loss on its loans to the private sector or its holdings of long-term assets, real resources in the form of Treasury tax revenues might be necessary to recapitalize the Fed, again to prevent inflation.

As noted earlier, the Fed has made a profit rather than a loss on its emergency lending. However, Bernanke (2013) notes that recent interest rate forecasts made by the CBO, Blue Chip consensus forecast, Survey of Professional Forecasters, and the Fed's interest rate models all suggested long-term rates could rise in the next few years. Analyses by Carpenter et al. (2013), Greenlaw et al. (2013), and Hall and Reis (2013) all note that if that were to happen, the Fed would realize substantial losses on its current holdings of MBS and long-term treasuries. On the other hand, the Fed also has significant unrealized capital gains on its gold holdings, currently valued at $42.22 an ounce. In the calculations for this project, I have decided to treat currency held by the public as entailing zero net off-balance-sheet liabilities for the Treasury or the Federal Reserve, meaning that I regard a standard open-market purchase of a Treasury security that ends up as more currency held by the public as reducing outstanding federal indebtedness by the amount of the purchase.

However, currency is no longer the most important entry on the liability side of the Fed's balance sheet. Figure 11 plots total Fed liabilities. Note that the height of that graph is by definition every week exactly equal to the height of the graph of total assets in Figure 10. As of April 3, 2013, currency held by the public came to $1,180 billion, whereas reserves were $1,838 billion, and all other liabilities $242 billion. Most of the reserves that the Fed created to

Figure 11
All Federal Reserve Liabilities
(seasonally unadjusted, from January 1, 2007, to April 3, 2013)



Source: Federal Reserve, "Factors Affecting Reserve Balances," *H41 Release*, 2007–2013.
Note: *Currency:* Currency in circulation. *Misc.:* Sum of Treasury cash holdings, foreign official accounts, and other deposits. *Other:* Other liabilities and capital. *Reserves:* Reserve balances with Federal Reserve Banks. *Reverse RP:* Reverse repurchase agreements. *Treasury:* Sum of U.S. Treasury general and supplementary funding accounts.

pay for its large-scale asset purchases are being held idle in banks' accounts with the Fed at the end of each day.

One key reason why banks are content to hold that huge new volume of reserves is that the deposits now earn interest. Granted, the rate is quite low: 25 basis points, or a 0.25 percent annual rate. However, with T-bills only paying 7 basis points as of April 2013, reserves are a more attractive asset than T-bills. As interest rates rise, the Fed will have to pay a higher interest rate in order to persuade banks to continue to hold the deposits with the Fed overnight. Purchasing a 10-year Treasury bond with newly created interest-bearing reserves is thus just a swap of one form of government debt

(a 10-year bond) for another (in effect, an overnight interest-bearing loan from the bank to the Fed). If we are going to subtract the Fed's holdings of T-bonds from net government indebtedness, we need to add reserve deposits with the Fed as an alternative new liability of the combined Treasury-Fed balance sheet.

Lines 4–7 of Table 3 show how the actions of the Federal Reserve change the on- and off-balance-sheet liabilities of the federal government. In line 4, we add reserve balances held by member banks. Although that is an unambiguous liability of the U.S. government, it is appropriate to treat it as an off-balance-sheet item because it is matched by corresponding assets. We then subtract that part of Treasury debt that is held by the Federal Reserve (line 5), as well as subtract agency debt and MBS purchased by the Federal Reserve (lines 6 and 7). We do so because we have already been counting the latter as part of our off-balance-sheet total, and thus view the Fed's purchase of MBS as swapping one off-balance-sheet liability (the government's implicit MBS guarantee) for another (interest-bearing reserves created by the Federal Reserve). If, however, one had not been persuaded that the MBS were already a federal liability, the Fed's actions should be viewed as making them more so, insofar as interest-bearing reserves are a more direct federal liability than the mortgage guarantees.

The bottom line is that the Federal Reserve's net contribution to off-balance-sheet liabilities (line 8) changed from −$773 billion in 2006 to +$360 billion at the end of 2008, a net swing of $1,133 billion. In other words, the net effect of the Fed's emergency lending between 2006 and 2008 was to increase the net indebtedness of the federal government by over $1 trillion, balanced by acquisition of corresponding assets (the emergency loans). As emergency lending was phased out, the Fed has returned to a position of on-balance-sheet liabilities reducing total government liabilities by $1,128 billion. Although the Fed's assets are greatly expanded relative to their historical levels, most of those new assets were already on- or off-balance-sheet liabilities of the broader federal government, so swapping them for reserves left total off-balance-sheet federal liabilities unchanged.

3.5 Federal Government Trust Funds

In addition to the $11,582 billion in Treasury debt held by the public at the end of calendar year 2012, the Treasury had issued an

additional $4,851 billion in obligations that were regarded as an asset held by other federal government entities. By far the most important of those obligations is the Social Security Trust Fund, which held $2,610 billion, a little more than half of the intragovernmental total. That sum is counted as a liability by the Treasury and as an asset by the Social Security Trust Fund. One perspective is that the sum is, therefore, money that the government owes to itself and, as a consequence, should not be included in a sensible accounting of net government indebtedness. But if we are to take the entirely reasonable position that the $2.6 trillion in Treasury securities held by the Social Security Trust Fund should not be counted as a net federal liability, we are forced also to regard the Social Security Trust Fund as holding no assets. However, there is a reason that Social Security wants to regard those Treasury obligations as an asset: the program anticipates significant liabilities associated with payments expected by current and future retirees.

It is true that those liabilities do not rise to the status of the "full faith and credit of the United States." The federal government might well choose to reduce payments to beneficiaries relative to those anticipated under the program's current practice, or else it might increase future payroll taxes. However, those are the same options that the government would consider in figuring out how to honor its official on-balance-sheet liabilities. The political difficulties that the government might face in making changes to the public's perceived Social Security obligations should reasonably be regarded as an important influence on the government's ability to honor its on-balance-sheet liabilities. For that reason, it seems entirely appropriate to include those implicit commitments in an accounting of the federal government's combined off-balance-sheet liabilities.

The Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds makes several efforts to estimate the present value of those obligations along with their offsetting tax receipts. Obviously, a lot of guesswork goes into such calculations, and the results can be extremely sensitive to assumptions. For example, if one were to use an interest rate for discounting that is less than the assumed growth rate, the concept of present value calculated over an infinite horizon does not even exist. For an interest rate only slightly bigger than the growth rate, small changes can produce big effects in long-horizon calculations.

Nevertheless, for a sense of scale if nothing else, I use one of the present-value concepts that has been used in both the Social Security and Medicare annual reports, which is to calculate the present value of future benefits to be paid to all current participants (that is, all Americans currently ages 15 or older). The annual reports also calculate the present value of future Social Security taxes to be paid by that same group. As of the end of calendar year 2012, the difference between the two—or the present value of the unfunded obligation for current participants—came to $26.5 trillion, up from $16.5 trillion in 2006 (see line 1 of Table 4).

Note that the numbers reported in Tables 1 and 3 refer to gross off-balance-sheet liabilities with no effort made to subtract offsetting assets. By contrast, the $26.5 trillion in Table 4 already subtracts the off-balance-sheet asset of Social Security represented by future tax revenues targeted for the program. However, insofar as these tax revenues are indeed used for this purpose, they will not be available for servicing the costs of the $11.3 trillion on-balance-sheet debt. For that reason, a case could be made for using the gross Social Security liability of $52.0 trillion, rather than the net liability of $26.5 trillion reported in Table 4, in our tabulation of total off-balance-sheet federal liabilities.

These numbers are so huge that it is hard to discuss them in a coherent way. As noted earlier, the calculations that go into them are easily challenged. But although one can quarrel with the specific numbers, they reflect an undeniably important reality: the U.S. population is aging, and an aging population means fewer people paying in and more people expecting benefits. That reality is unambiguously going to be a key constraint on the sustainability of fiscal policy for the United States. One would think that we as a nation should be saving today in preparation for retirement. If, in fact, we are not, the currently enormous on-balance-sheet federal debt is all the more of a concern.

Similar calculations are reported by the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. Medicare Part A (hospital insurance) reported a present value of unfunded obligations for current program participants of $14.5 trillion as of the end of 2008. The Health Care and Education Reconciliation Act of 2010 added new revenues and reduced the Trustees' anticipated costs so that the unfunded

Table 4

Contribution of Government Trust Funds to Off-Balance-Sheet Federal Liabilities (as of end of indicated calendar year, in billions of dollars)

	2006	2008	2010	2012
1. Social Security	$16,500	$18,700	$21,400	$26,500
2. Medicare Part A (hospital insurance)	$11,800	$14,500	$8,000	$9,600
3. Medicare Part B (medical insurance)	$10,400	$13,500	$11,500	$13,100
4. Medicare Part D (prescription drug insurance)	$6,300	$5,200	$5,400	$4,900
5. Sum of Medicare liabilities	**$28,500**	**$33,200**	**$24,900**	**$27,600**
6. Other government trust funds	$1,308	$1,487	$1,646	$1,862
7. Combined contribution of government trust funds	**$46,308**	**$53,387**	**$47,946**	**$55,962**

Sources: **Line 1:** Present value of future Old Age, Survivor, and Disability Insurance cost less future taxes for current participants. From Table IV.B7, "Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Funds," 2007, 2009, 2011, and 2013. **Line 2:** Present value of future expenditures less income for current participants in Medicare Part A. From Table III.B11 or V.G2, "Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds," 2007, 2009, 2011, and 2013. **Line 3:** Present value of future contributions for current participants in Medicare Part B coming from general federal revenues. From Tables III.C16, III.C12, or V.G4, "Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds," 2007, 2009, 2011, and 2013. **Line 4:** Present value of future contributions for current participants in Medicare Part D coming from general federal revenues. From Tables III.C24, III.D8, or V.G6, "Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds," 2007, 2009, 2011, and 2013. **Line 5:** Sum of lines 2–4. **Line 6:** Government account series held by Airport and Airway Trust Fund, Employees Life Insurance Fund, Exchange Stabilization Fund, federal employees retirement funds, FSLC Resolution Fund, Highway Trust Fund, National Service Life Insurance Fund, Postal Service Fund, Railroad Retirement Account, Unemployment Trust Fund, and other. From Table FD-3, *Treasury Bulletin*, December 2008 and December 2012. **Line 7:** Sum of lines 1, 5, and 6. Note: Social Security and Medicare entries represent values as of end of calendar year, except the last entry, which is the end of calendar year 2011, the most recently available as of the time this article was written. Other government trust funds represent values as of end of indicated fiscal year.

obligation fell to $8.0 trillion by the end of 2010, but it subsequently rose to $9.6 trillion by the end of 2012 (see line 2 of Table 4). Medicare Part B (medical insurance) added an additional unfunded liability of $13.1 trillion as of the end of 2012, while Part D (prescription drug insurance) added another $4.9 trillion, for a total of $27.6 trillion in unfunded obligations currently reported for Medicare.

Again, these numbers represent the *net* off-balance-sheet liabilities associated with Medicare. For the gross liabilities (that is, without subtracting anticipated tax revenue), we would add $4.8 trillion to the total for Medicare Part B and $1.1 trillion for Part D. The Trustees Report does not itemize separately the numbers behind the net calculations for Part A, but it is clear that the upward adjustment necessary to arrive at a gross figure would be quite substantial. For example, in 2013 alone, Part A is expecting to collect $231.2 billion from payroll taxes and taxation of Hospital Insurance Benefits. By 2022, the annual number is expected to grow to $410.8 billion (see Table III.B4 of the Trustees Report). A projection of these future revenues has already been subtracted from the future outlays in arriving at the $9.6 trillion unfunded liability for Part A reported in Table 4.

A number of other government trust funds, like Social Security and Medicare, also claim as their assets debt obligations of the U.S. Treasury that are not included in the $11.3 trillion in Treasury debt held by the public. Like the Social Security Trust Fund, they are neither an asset nor a liability of a unified federal balance sheet, but they are simply an "IOU" from one arm of the government to another. But like Social Security and Medicare, although the government has no assets backing these funds, they may entail significant future fiscal obligations.

The largest of these funds are the Civil Service Retirement and Disability Fund, which claimed $904 billion in Treasury obligations among its assets as of the end of calendar year 2012, and the Military Retirement Fund, to which the Treasury owes $429 billion. Rather than try to go through each of the programs and assess the implicit or explicit commitments the government has made through them, I have taken the conservative approach (and I suspect, a recklessly conservative approach) of assuming that the programs are actuarially balanced—that is, assuming that the present value of commitments associated with the Civil Service Retirement and Disability Fund is exactly $904 billion. Since there are no assets to back

these commitments, future tax increases or spending cuts will be necessary to honor them, so I count this trust fund as adding $904 billion to the 2012 off-balance-sheet federal liabilities. Adding together all the government trust funds not treated elsewhere in this analysis, I calculate the contribution of these commitments to total off-balance-sheet liabilities to have been $1,862 billion as of 2012. The combined contribution of Social Security, Medicare, and other trust funds to the total off-balance-sheet federal liabilities is $55,962 billion, as shown in Table 4.

3.6 Other Off-Balance-Sheet Liabilities

A number of other off-balance-sheet commitments are potentially substantial, but they are not included in the totals given earlier. One that could prove to be quite important is the Pension Benefit Guaranty Corporation (PBGC), an independent agency of the U.S. government that was created by the Employee Retirement Income Security Act of 1974. The PBGC guarantees basic pension benefits for about 43 million Americans participating in 26,000 private-sector defined-benefit pension plans. The potential notional liabilities are enormous, but they are very difficult to assign a dollar value to. The federal government has never declared the guarantees to be backed by the full faith and credit of the U.S. government, and unlike Fannie and Freddie, the PBGC is not in conservatorship. For these reasons, I have not attempted to include pension guarantees as a current off-balance-sheet liability of the U.S. government. But that omission does not mean that pension benefits will not prove to develop into a very important fiscal challenge at the federal, state, and local government levels.

Another important commitment involves veterans' benefits. A 2010 assessment by the CBO estimates that by 2020, the annual cost of providing health care services to veterans who seek benefits through the Department of Veterans Affairs could be $69 billion to $85 billion, which would represent increases of 45–75 percent over 2010 levels.

Federal flood insurance is another potentially important off-balance-sheet federal liability. King (2013) estimates that government payouts for Hurricane Sandy under the National Flood Insurance Program could be $12 billion to $15 billion, in comparison with the $4 billion cash assets of the program. A study by the CBO (2007) estimates that as of February 2007, the 5.4 million policies issued through

the Federal Emergency Management Agency had a total coverage of $1 trillion. Again, I have made no effort to include the potential liabilities of these and other programs in the totals reported in this paper.

4. HOW CONCERNED SHOULD WE BE?

Table 5 summarizes the contributions of the various components surveyed in Section 3. I calculate total off-balance-sheet federal liabilities to have been $70.1 trillion as of 2012, six times the size of the federal debt itself. That total comes from a range of different programs, each of which is associated with its own benefits and its own concerns. In the case of the FDIC guarantees and emergency Fed lending, they seemed to accomplish what they were intended and in my opinion do not pose significant risks to taxpayers in the current environment. Other programs, such as the federal government's big role in lending for housing and education, have less clear benefits and have been associated with more tangible costs. The biggest off-balance-sheet liabilities come from recognition of the fiscal stress that will come in the form of an aging population and rising medical expenditures.

Table 5
Treasury Debt Held by the Public and Combined Federal Off-Balance-Sheet Liabilities (in billions of dollars)

	2006	2008	2010	2012
1. Treasury debt held by public	$4,867	$5,837	$9,052	$11,299
2. Housing-related commitments	$6,386	$8,036	$7,594	$7,520
3. Student and other loan guarantees	$468	$547	$419	$325
4. FDIC	$4,154	$4,975	$6,575	$7,406
5. Federal Reserve	−$773	$360	−$1,136	−$1,128
6. Social Security	$16,500	$18,700	$21,400	$26,500
7. Medicare	$28,500	$33,200	$24,900	$27,600
8. Other government trust funds	$1,308	$1,487	$1,646	$1,862
9. Total off-balance-sheet commitments	**$56,544**	**$67,305**	**$61,398**	**$70,085**

Sources: **Line 1:** From line 1 of Table 1. **Line 2:** From line 4 of Table 1. **Line 3:** Sum of lines 5–7 of Table 1. **Line 4:** From line 3 of Table 3. **Line 5:** From line 8 of Table 3. **Line 6:** From line 1 of Table 4. **Line 7:** From line 5 of Table 4. **Line 8:** From line 6 of Table 4. Sum of lines 2–8 in Table 5.

It is worth noting that there are many historical episodes in which off-balance-sheet liabilities ended up having quite significant on-balance-sheet implications. One example is provided by the problems with savings and loans in the 1980s. Losses at those institutions ended up dwarfing the capabilities of the now-defunct Federal Savings and Loan Insurance Corporation to honor its promise to guarantee depositors. Curry and Shibut (2000) estimate that the final on-balance-sheet cost to U.S. taxpayers of honoring those off-balance-sheet guarantees comes to $124 billion.

Some economists see the Asian crisis of 1997 as providing a more dramatic illustration. A number of Asian countries experienced rapid and dramatic currency depreciation and spiking interest rates that led to significant drops in real economic activity. According to Burnside, Eichenbaum, and Rebelo (2001), "A principal cause of the 1997 Asian currency crisis was large prospective deficits associated with implicit bailout guarantees to failing banking systems." The authors note, for example, that the costs of restructuring and recapitalizing the banking system in those episodes amounted to 65 percent of GDP for Indonesia, 24 percent of GDP for South Korea, 22 percent for Malaysia, and 35 percent for Thailand. The authors suggest that market perceptions of the prospective deficits associated with bailing out problem banks were the trigger for the rapid capital flight from those countries.

More recently, the Irish government entered the Great Recession in a seemingly very strong fiscal position, with gross government debt totaling only 25 percent of GDP. But on September 30, 2008, in response to developing strains on the financial system, the government guaranteed all deposits, covered bonds, senior debt, and dated subordinated debt of the country's six largest banks. By July 2011, Ireland needed to inject 64 billion euros (45 percent of 2010 GDP) into the banking system. As a result, the Irish government budget deficit for 2010 amounted to 31 percent of the country's GDP, and its debt level rose above 100 percent of GDP. The interest rate on 10-year Irish government bonds went from 4.2 percent in 2008 to 14 percent in July 2011 as investors fled Irish sovereign debt.

I am not predicting that a similar crisis is on the verge of unfolding for the United States. Some may argue that the current off-balance-sheet liabilities of the U.S. federal government are smaller than those tabulated here; others could arrive at larger numbers. These off-balance-sheet concerns may or may not translate into

significant on-balance-sheet problems. But one thing seems undeniable: they are huge. And implicit or explicit commitments of such a huge size have the potential to have huge economic consequences, perhaps for the better, perhaps for the worse. Acknowledging their size is a necessary first step for making wise policy decisions.

REFERENCES

Ashcraft, Adam B., and Til Schuermann. 2007. "Understanding the Securitization of Subprime Mortgage Credit." Working paper, Federal Reserve Bank of New York.

Bagehot, Walter. 1873. *Lombard Street: A Description of the Money Market.* London: H. S. King.

Bencivenga, Valerie R., and Bruce D. Smith. 1993. "Some Consequences of Credit Rationing in an Endogenous Growth Model." *Journal of Economic Dynamics and Control* 17: 97–122.

Bernanke, Ben S. 2013. "Long-Term Interest Rates." Speech at the Annual Monetary/Macroeconomics Conference, San Francisco, March 1, http://www.federalreserve.gov/newsevents/speech/bernanke20130301a.htm.

Bhattacharya, Sudipto, Arnoud W. A. Boot, and Anjan V. Thakor. 1998. "The Economics of Bank Regulation." *Journal of Money, Credit, and Banking* 30: 745–70.

Brunnermeier, Markus K., and Yuliy Sannikov. 2012. "A Macroeconomic Model with a Financial Sector." Working paper, Princeton University, Princeton, NJ.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo. 2001. "Prospective Deficits and the Asian Currency Crisis." *Journal of Political Economy* 109: 1155–97.

Carpenter, Seth B., Jane E. Ihrig, Elizabeth C. Klee, Alexander H. Boote, and Daniel W. Quinn. 2013. "The Federal Reserve's Balance Sheet: A Primer and Projections." Working paper, Federal Reserve, Washington, DC, http://www.federalreserve.gov/pubs/feds/2013/201301/201301pap.pdf.

Chaney, Paul K., and Anjan V. Thakor. 1985. "Incentive Effects of Benevolent Intervention: The Case of Government Loan Guarantees." *Journal of Public Economics* 26: 169–89.

Congressional Budget Office. 2001. "Federal Subsidies and the Housing GSEs." Washington, DC, http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/28xx/doc2841/gses.pdf.

_____. 2007. "Value of Properties in the National Flood Insurance Program." Washington, DC, http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/82xx/doc8256/06-25-floodinsurance.pdf.

_____. 2010. "Potential Costs of Veterans' Health Care." Washington, DC, http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/118xx/doc11811/2010_10_7_vahealthcare.pdf.

Cooper, Russell, and Thomas W. Ross. 2002. "Bank Runs: Deposit Insurance and Capital Requirements." *International Economic Review* 43 (1): 55–72.

Curry, Timothy, and Lynn Shibut. 2000. "The Cost of the Savings and Loan Crisis: Truth and Consequences." *FDIC Banking Review* 13: 26–35.

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

Diamond, Douglas W., and Raghuram G. Rajan. 2000. "A Theory of Bank Capital." *Journal of Finance* 55 (6): 2431–64.

Flannery, Mark J. 1996. "Financial Crises, Payment System Problems, and Discount Window Lending." *Journal of Money, Credit and Banking* 28 (4, Part 2): 804–24.

Gokhale, Jagdeesh. 2013. *Spending beyond Our Means: How We Are Bankrupting Future Generations.* Washington, DC: Cato Institute.

Greenlaw, David, James D. Hamilton, Peter Hooper, and Frederic S. Mishkin. 2013. "Crunch Time: Fiscal Crises and the Role of Monetary Policy." U.S. Monetary Policy Forum, New York, February 22 (revised July 29), http://dss.ucsd.edu/~jhamilto/USMPF13_final.pdf.

Haffner, Marietta E. A. 2008. "Subsidization as Motor to Residential Mortgage Securitization in the U.S." *Journal of Housing and the Built Environment* 23: 337–51.

Hall, Robert E., and Ricardo Reis. 2013. "Maintaining Central-Bank Solvency under New-Style Central Banking," Working paper, Stanford University, Stanford, CA.

Hamilton, James D. 2007a. "Commentary: Housing and the Monetary Transmission Mechanism." In *Housing, Housing Finance, and Monetary Policy*, pp. 415–22. Kansas City, MO: Federal Reserve Bank of Kansas City.

_____. 2007b. "Fannie, Freddie, and Ben." *Econbrowser*, March 11, 2007, http://www.econbrowser.com/archives/2007/03/ fannie_freddie.html.

_____. 2012. "U.S. Government Profits from AIG Bailout." *Econbrowser*, December 13, http://www.econbrowser.com/archives/2012/12/ us_government_p.html.

Hamilton, James D., and Jing Cynthia Wu. 2012. "The Effectiveness of Alternative Monetary Policy Tools in a Zero Lower Bound Environment." *Journal of Money, Credit and Banking* 44 (1, Supplement): 3–46.

Judson, Ruth. 2012. "Crisis and Calm: Demand for U.S. Currency at Home and Abroad from the Fall of the Berlin Wall to 2011." Working paper, Federal Reserve, Washington, DC.

Kareken, John H., and Neil Wallace. 1978. "Deposit Insurance and Bank Regulation: A Partial-Equilibrium Exposition." *Journal of Business* 51 (3): 413–38.

King, Rawle O. 2013. "The National Flood Insurance Program: Status and Remaining Issues for Congress." Congressional Research Service, Washington, DC, February 6, http://www.fas.org/sgp/crs/misc/R42850.pdf.

Lee, Donghoon. 2013. "Household Debt and Credit: Student Debt." Slide presentation, Federal Reserve Bank of New York, February 28, http://www.newyorkfed.org/newsevents/mediaadvisory/2013/Lee022813.pdf.

Santos, Joáo A. C. 2001. "Bank Capital Regulation in Contemporary Banking Theory: A Review of the Literature." *Financial Markets, Institutions and Instruments* 10 (2): 41–84.

Shiller, Robert J. 2000. *Irrational Exuberance*. Princeton, NJ: Princeton University Press.

Smith, Bruce D., and Michael J. Stutzer. 1989. "Credit Rationing and Government Loan Programs: A Welfare Analysis." *American Real Estate and Urban Economics Association Journal* 17 (2): 177–93.

U.S. Department of Housing and Urban Development. 2012. "Annual Report to Congress: Fiscal Year 2012 Financial Status of FHA Mutual Mortgage Insurance Fund." Washington, DC.

Comment

Douglas Holtz-Eakin

James Hamilton has provided an outstanding overview of the level and recent growth of federal off-balance-sheet (OBS) liabilities. His striking conclusion is that those obligations exceed $70 trillion, nearly six times as much as the official federal debt outstanding. In addition to tabulating the size of those liabilities, Hamilton examines the conventional reasons for their existence and analyzes the threat they pose to the United States. It is an extremely valuable, timely overview.

My comments will focus on issues of valuation, the timing of those obligations, and their economic significance. In general, the OBS liabilities pose some difficult issues in valuation. To start, notice that under current law Social Security benefits will be cut roughly 25 percent across the board when the trust fund exhausts (currently estimated to occur in 2033). At that point, benefits are to be reduced to match inflows of payroll taxes; as a result, no new net liability is created.

Alternatively, one could argue that the current benefits *policy* represents the commitment. If so, future benefits will exceed future payroll taxes by roughly 2 percent of gross domestic product in perpetuity, with a correspondingly large off-balance-sheet liability. If, as Hamilton discusses, the OBS liabilities are a measure of commitment to retirees, the latter may be a better measure.

Medicare (and Medicaid) presents a different issue. Medicare spending is projected to rise at annual rates of 5–7 percent. At present, that projection exceeds any plausible discount rate that might

Douglas Holtz-Eakin is president of the American Action Forum and a former director of the Congressional Budget Office.

be applied to OBS liability growth.[1] Mechanically, that means that the scale of those liabilities is *infinite*. To compute a present value, most analysts assume that the growth in health care cost will slow. However, the *date* at which the slowing begins, the *pace* at which it ultimately grows, and the *long-run growth rate* are often quite arbitrary decisions. At the same time, those decisions fundamentally determine the scale of the OBS liability.

Housing programs present slightly different valuation problems. For years leading up to the bursting of the housing bubble and financial crisis, it was widely recognized that Fannie Mae and Freddie Mac benefited from the perception of implicit taxpayer backing. One way to estimate the value provided is to view the taxpayer subsidy as an implicit option for Fannie Mae and Freddie Mac to "put" their debt obligations in the event of financial distress.

From that perspective, the OBS liability is simply the value of that put option.[2] *Ex post*, the value is quite different, as the federal government has assumed the obligation to back nearly $5 trillion in debt outstanding for the housing government-sponsored enterprises.

The upshot is that the reader should pay a great deal of attention to the overall scale of OBS liabilities. But the degree of precision is necessarily limited, and particular values should be taken with a grain of salt.

The second issue related to valuation is the pace at which the liabilities accumulate. Present-value calculations (such as those presented by Hamilton) are designed to eliminate matters of timing—a present-value dollar in 2013 is treated the same as a present-value dollar in 2033. However, the reality of the budget process is that

[1] Recent data have shown an overall slowing in the growth of national health care spending, raising the possibility that this problem could be over. There is reason to be skeptical of this seemingly good news, however. First, the Great Recession dampened the growth of prices and spending across the economy; health care is no different. As the economy recovers, one would expect the pace to pick back up. Second, many times in the past, health care spending grew at the pace of gross domestic product or slower—exactly what is happening right now—notably in the late 1990s. In each case, the pace turned right around and picked back up. Finally, as the Patient Protection and Affordable Care Act (more commonly known as Obamacare) is implemented, it will cover millions more people with health insurance precisely for the purpose of making sure they spend more on health care—a recipe to see current trends reversed.

[2] See Congressional Budget Office, "Updated Estimates of the Subsidies to the Housing GSEs," April 8, 2004, http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/53xx/doc5368/04-08-gse.pdf.

timing does matter. As the baby boom generation retires, many of the OBS liabilities will be transformed to on-budget demands for taxpayer resources. That transformation has been widely anticipated for quite some time. However, it was not anticipated that the transformation would occur in the immediate aftermath of a large run-up in the other OBS liabilities.

That brings me to my final point: how dangerous are the OBS liabilities? Hamilton expresses concern, but I see the fiscal situation as much more dangerous. Given the high levels of U.S. official debt, any transformation of OBS liabilities into additional debt will guarantee that debt service will eliminate budgetary flexibility. In addition, the cash-flow deficits engendered would require substantial increases in the level of taxation, a move that would suppress growth and lower standards of living. And even with higher levels of taxation, the United States would face a substantial probability of the loss of international investor confidence and a sovereign debt crisis. In short, modifying spending programs to reduce the scale of OBS liabilities should be at the top of any policymaker's to-do list.

To summarize, Hamilton has provided a valuable insight into an important fiscal threat. It is a must-read for the policy community.

Comment

Phillip Swagel

The United States' long-term fiscal challenge is the product of government promises that represent enormous implicit liabilities. A response to that challenge requires making difficult policy choices. Understanding the magnitude of those liabilities is a vital first step—a prerequisite, really—for coming up with good policies. James Hamilton, in his paper, provides a valuable public service by assembling a comprehensive listing and discussion of the $70 trillion of U.S. implicit liabilities.

There are a limited number of fundamental policy choices that can be used to resolve the U.S. fiscal imbalance—even while there are myriad fine choices within broad policy categories. Broadly speaking, addressing the long-term fiscal challenge involves some combination of

- increased revenue from taxes within an entitlement program, such as Social Security or Medicare;
- increased revenue from other sources, such as income taxes;
- decreased spending within a given program that has implicit liabilities; and
- decreased spending elsewhere in the government sector to free up resources.

More simply, the choices are to adjust revenues or spending, whether inside or outside the particular programs with the implicit liabilities. Steps meant to increase macroeconomic growth fit within those choices in the sense that stronger growth would lead to increased revenues and perhaps changes in spending, such as lower transfer payments. In any case, the actuarial reports for Social

Phillip Swagel is a professor of public policy at the University of Maryland. He was assistant secretary for economic policy at the U.S. Department of the Treasury from 2006 to 2009.

Security and Medicare make clear that even an optimistic increase in economic growth is not likely to stabilize those programs' finances. There is no "third way" option that avoids difficult choices regarding spending and revenues.

Calculating the magnitude of those obligations as Hamilton has done makes clear that the choices involved are neither modest nor easy. If there were simple or painless choices involved, they would have been made already. Similar considerations hold for other implicit liabilities, such as implicit government guarantees.

Judgments regarding who should bear the adjustments' burden ultimately will guide the policy choices, regardless of whether they lean toward decreased spending or increased revenue. In the context of entitlements such as Medicare and Social Security, someone in the future will either receive fewer benefits than currently promised or face higher taxes than today. This choice can be seen as a counterpart to the question of who will or should benefit from the obligations that correspond to the implicit liabilities. The answer to each has important implications for the potential policy options that are available to address the liabilities. My comments on Hamilton's paper focus on how the nature and magnitude of the implicit liabilities influence the policy choices.

(Lack of) Transparency—A Common Motivation

Implicit liabilities arise for many reasons, including (a) to prevent or respond to crises, such as financial panics; (b) to subsidize targeted activities, such as affordable housing and college loans; and (c) to provide retirement security or health care to retirees. Those reasons can be seen as laudable goals.

A common factor, however, in taking on implicit rather than explicit liabilities is the desire to avoid transparency for the obligation. In the recent financial crisis, for instance, the Federal Reserve's balance sheet became a part of the financial rescue through the use of nonrecourse lending. The Fed lent money to entities against specified collateral, but without the ability to call on other assets of the borrower if losses exceeded the value of the collateral. In effect, the U.S. government owned at least the downside of the risky assets involved. This device was used in a variety of ways, including for the rescues of Bear Stearns, AIG, the Commercial Paper Funding Facility, and the Term Asset-Backed Securities Loan Facility. In each

of those cases, the Fed provided liquidity to help stabilize firms it considered to be systemically important or else to unfreeze particular credit channels. But the Fed did that in a way that involved a greater risk of loss than is typical with its lending activities. The key reason for using nonrecourse lending was that doing so was politically possible, while using other government resources generally would have required an act of Congress (in the case of Bear Stearns or AIG, both of which went into distress before the advent of the Troubled Asset Relief Program [TARP]) or would have meant drawing heavily on the limited amount of funds that could be obligated through TARP.

As a case in point, consider the Fed involvement with the J. P. Morgan–Bear Stearns transaction. Its involvement was born out of necessity, since policymakers felt that it was essential to prevent a collapse of Bear Stearns on account of its position in the triparty repo system (this position seemed reasonable at the time, but it remains a point of controversy). The Fed lent money to a limited liability corporation that, in turn, purchased $30 billion of securities that J. P. Morgan did not want to accept as part of its acquisition of Bear Stearns' assets. J. P. Morgan agreed to cover the first $1 billion of the loss, meaning that the Fed was exposed to $29 billion of risk if the value of the securities used as collateral went to zero. In the end, the Fed was paid back (and then some) for its loans, but that success was far from certain in March 2008 when the loans were first made.

By structuring the activities as an implicit guarantee through nonrecourse lending, the Fed's involvement was nontransparent, particularly to members of Congress. Lawmakers did not immediately understand that the Fed had taken on the downside of $29 billion of assets that were too risky for J. P. Morgan. By the fall of 2008, so many interventions were structured in this way that the continued use of nonrecourse financing no longer seemed unusual, even though each instance exposed taxpayers to loss and, therefore, could be seen as an implicit liability (along with the asset of the collateral).

In the policy areas of education and housing finance, the lack of transparency is with regard to the accounting treatment of the federal liabilities. Student loans and guarantees for mortgages provided by the Federal Housing Administration (FHA) are scored on the federal budget using the accounting treatment specified in the Fair Credit Reform Act (FCRA), which considerably underestimates the risk being taken on by the federal government. The FCRA accounting

treatment sometimes allows a financial activity to show a profit for the federal government when a private company would book a loss. It is hard to believe that the federal government is intrinsically better at a financial activity than the private sector; if anything, the opposite is more likely the case. Any claim of profitability of federal guarantees on student loans and of FHA guarantees on new mortgages could be a figment of the accounting—indeed, that is the calculation of the Congressional Budget Office when restating the federal exposure using the more appropriate standard of fair value accounting. The eventual losses, however, will be all too real.

Use of fair value accounting provides a more accurate assessment of risk and thus of the financial implications of implicit liabilities. But it elicits considerable opposition. In the context of housing, for example, proponents of an expansive role for the federal government argue against fair value accounting, presumably out of concern that writing down the full cost of the federal government's support for housing would lead to less political and popular support for these programs. In the case of student loans, the federal takeover of student lending was justified as a way for the federal government to make money—again, because of the understatement of risks involved with the Fair Credit Reform Act's accounting treatment. It will not be a surprise in the future when the bill comes due.

Similar considerations can be seen with implicit pension liabilities, where the lack of transparency is a political boon—a way to spend money without transparency. The future pension liabilities eventually become those of the present, as promises made to current workers turn into checks to be written to retirees. Still, the temptation to turn to implicit liabilities is clear.

At the federal level, entitlement reform policy has similar elements of nontransparency as a virtue (using that word loosely and with some distaste). In Social Security, for example, the focus on a 75-year horizon over which to assess the actuarial condition of the program's finances gives rise to "solutions" that bring in revenues up front but then result in expenditures down the road. This distorted view takes advantage of the structure of Social Security funding in which individuals pay taxes while working and then draw contributions in retirement. Using a limited horizon—even a long one such as 75 years—skews the analysis of ways to solve the implicit liabilities because policies with up-front revenues will

Table 1
Impact on Social Security Actuarial Balance as a Percentage of Taxable Payroll over 75 Years and in the 75th Year

Provision	Change over 75 Years	Change in 75th Year Only
Chain-weighted by Consumer Price Index	0.54	0.73
Progressive indexation of benefits	1.43	4.10
Cover state and local employees	0.16	−0.17
Increase retirement age slowly from 67 to 68	0.32	0.72
Increase retirement age quickly from 67 to 68	0.42	0.72
Tax all earnings, and allow benefits to accrue	1.92	1.64
Tax all earnings but without benefit accrual	2.36	2.49
Balance 2012 trustees' intermediate assumptions	−2.67	−4.50

appear to make more progress in addressing the fiscal imbalance than is actually the case.

This phenomenon can be seen by examining the effect of various policy proposals in improving the financial condition of Social Security, thereby comparing the average impact over 75 years with the improvement in the 75th year alone. Table 1 shows the improvement in the financial position of Social Security expressed in percentage points of the taxable wage base. The table also shows several policy proposals. The first column shows the effect on the total imbalance over 75 years, while the second column shows the impact in the 75th year only. A policy with a greater impact on average over the 75 years than in the 75th year will be one with the property noted earlier, thus overstating the contribution toward sustainable solvency of Social Security by limiting the time horizon considered.

The first two proposals involve slowing the growth of benefits by changing the price index used to calculate benefits changes over time. The chain-weighted Consumer Price Index would take into account the substitution across consumption items to be expected in response to price changes, while the progressive indexation would result in slower benefit growth for people with high lifetime incomes than for people with low lifetime incomes. As the table shows, both policy options have the characteristic that the improvement in Social

Security finances grows over time, as the correction from the index change translates into persistently slower expenditure growth. Retirees into the future will pay the same amount, but they will receive increasingly more modest benefits than those now promised (but which cannot be afforded under the current system).

A similar phenomenon can be seen for the proposals in the table to increase the retirement age from 67 to 68: the impact in narrowing the financial gap increases over time, with the progress in the 75th year greater than on average over the 75 years. An increase in the retirement age is akin to a reduction in benefits—in this case, for everyone. Note that the speed of the change to a higher retirement age affects the average improvement over 75 years but not in the 75th year, because by 75 years out both changes will have increased the retirement age by the full year. The fiscal improvement of the more rapid increase in the retirement age results from making more generations of retirees subject to a higher retirement age.

In contrast, the policy option to subject all earnings to the Social Security tax has the property that the fiscal improvement in the 75th year is smaller than the average over all 75 years. This fiscal improvement comes about because removing the taxable wage cap initially brings in new revenues, while the associated benefits are accrued but not yet paid out. As workers retire over time and collect the benefits associated with their higher contributions, however, the net improvement wanes. Note that the fiscal improvement does not shrink over time in the alternative shown in the table in which workers do not accrue benefits for their additional taxes above the current maximum, though this variant would break the link between contributions and benefits that is a hallmark of the Social Security program. If one considers options to shore up the finances of Social Security, revenues can be added or benefit growth slowed. But the intergenerational nature of the implicit liabilities in the system tends to skew the policy debate by making some options seem more effective over a limited horizon than they are when a longer one is considered.

Recent policy changes that purport to strengthen Medicare likewise take advantage of accounting limitations with regard to the tracking of the financial status of the program. Medicare spending was reduced as part of the Patient Protection and Affordable Care Act (PPACA), with the lower payments to Medicare being reallocated to fund the costs of expanding health insurance coverage.

The accounting for Medicare, however, looks only within the system and not at the external payments for PPACA. That approach means that within the narrow purview of Medicare finances, the system appears to be strengthened by payment reductions when, in fact, the economic resources that are notionally freed up are meant to be devoted in the legislation to providing new benefits under PPACA.

It is naturally tempting for policymakers to take advantage of this complication. The 2013 press statement from the Treasury Department made at the time of the release of the Medicare Trustees report does exactly that, asserting that PPACA has extended the life of the trust fund: "Medicare's Hospital Insurance (HI) Trust Fund is projected to exhaust its assets in 2026, two years later than was projected in last year's report, and nine years later than was projected in the last report released prior to passage of the Affordable Care Act."[1]

That assertion is misleading in that the economic resources are being used to pay for PPACA. The Trust Fund exhausts its assets later only because of the lack of transparency in the accounting for the implicit liabilities. Of course, that is what makes it so attractive from a political perspective to use cuts in Medicare to pay for new programs such as PPACA.

How Will These Implicit Liabilities Be Resolved?

The broader point of noting the nontransparency of implicit liabilities is that this characteristic makes them politically attractive as a financing source. The difficult choices are postponed while political credit is taken up front. Eventually, however, difficult choices must be made to deal with the fiscal imbalance.

For state-level liabilities, I believe that a relatively straightforward algorithm explains how critical implicit liabilities will be addressed: taxes will be increased as much as is politically feasible, and only then will spending be cut. This algorithm results from my observation that the largest state-level fiscal imbalances are in so-called blue states—those controlled by governors and legislators aligned with the Democratic Party. In those states, the political system is less averse to tax increases than spending cuts that hit government-aligned political constituencies.

[1] U.S. Department of the Treasury, "Statement by Secretary Lew on the Releases of Social Security and Medicare Trustees Reports," news release, May 31, 2013, http://www.treasury.gov/press-center/press-releases/Pages/jl1963.aspx.

California and Illinois provide cases in point. The fiscal imbalances in both states are chronic situations that only occasionally rise into the public debate. In Illinois, a modest change to public pensions was made in the fall of 2013, but only after large tax increases over several years. In California, on which Hamilton has written extensively, media reports say the fiscal situation has improved and is heading to surpluses, but this result obtains only because the figures ignore unfunded pension liabilities. Taxes were raised first, in the form of Proposition 30, which was advertised as providing new funding for education. However, in the end, much of the resources generated from the higher taxes will go to fund pensions. That change is not exactly what voters thought they were getting in voting for Proposition 30, but it is an inevitable consequence of the use of implicit liabilities to fund promises made in the past.

Broadly speaking, resolution of state-level implicit liabilities will result from a negotiation between governors, taxpayers, and interest groups, notably public-sector unions. The outcome will be to raise taxes as much as possible and then to cut spending as slowly as possible. Meanwhile, the composition of state spending will be affected by the drain from current revenues. In California, for example, the need to cover pension costs appears to affect a variety of social outcomes, including the quality of the education system through higher student-to-teacher ratios and crime rates through prisoner releases dictated by prison underfunding.

At the federal level, Hamilton's paper makes clear that dealing with the fiscal challenge requires addressing the Social Security and Medicare entitlements. That is the case even though the overall fiscal outlook has improved in the past year through (a) a combination of stronger economic growth and thus tax revenues, (b) higher tax rates as a result of the "fiscal cliff" tax deal that took hold in January 2013, and (c) lower spending brought about by the federal sequester. The Congressional Budget Office reports that the 75-year fiscal gap is down to 1.7 percent of U.S. gross domestic product—not a large enough amount to seem urgent or a threat to fiscal and macroeconomic stability.

The problem with this fiscal gap estimate is that it assumes that current law holds, even though a variety of spending cuts and tax increases embedded in current law are unlikely to be allowed to take place. For example, reductions to the payments physicians receive

for treating Medicare patients have been repeatedly averted, and a wide range of business tax cuts likewise have been repeatedly extended (indeed, they are known as the "tax extenders"). PPACA adds additional payment "cliffs," including substantial reductions in payments for Medicaid providers, such as doctors who treat Medicaid recipients. If the temporarily higher payments put into place with PPACA are not allowed to expire—and presumably there will be pressure to avoid payment reductions that could make it difficult for patients to receive care—then the actual fiscal imbalance is higher. In this sense, the lack of progress in addressing the fiscal imbalance represented by implicit liabilities can be seen as a political approach of "whistling past the graveyard."

The use of implicit liabilities affects policies in financial markets. The Federal Deposit Insurance Corporation (FDIC), for example, has the legal authority to borrow from the Treasury if the FDIC's deposit insurance fund is low, with the deficit to be made up in the future through higher insurance premiums. In practice, however, the FDIC works hard to avoid having to tap the Treasury, in large part because that would present a bureaucratic loss of face. Instead, the FDIC has taken other steps to shore up the deposit insurance fund, including having banks make advance payments on their insurance premiums, even while policymakers were concerned about the potentially adverse effects on lending from banks with inadequate capital.

Implicit liabilities are also accumulated through the orderly liquidation fund created in Title II of the Dodd-Frank financial regulatory reform legislation. That fund allows policymakers to put public money into a failing financial institution in order to continue its operations, while imposing any losses on shareholders and bondholders of the institutions and then ultimately on other financial firms, if needed, to ensure that taxpayers do not take losses. In contrast to the approach taken with the FDIC, a resolution fund is not built up in advance, but instead costs are imposed on industry participants, who should know in advance that they are at risk (for example, bondholders should know that they face the possibility of having haircuts imposed and should build that into the yields they require to lend to banks). This structure is reminiscent of the approach taken by the Federal Reserve in providing support for Bear Stearns and AIG, as discussed earlier.

The use of implicit liabilities influenced the policy choice for the government takeover of Fannie Mae and Freddie Mac. The agreements struck between the Treasury and each firm provided for taxpayer capital to be made available as needed to ensure solvency. Taxpayers were thus on the hook for any losses embedded in the government-sponsored enterprises. The obligation is explicit, but the amount of support is unknown and is not paid for in advance. The policy choice can thus be seen as involving an implicit liability, even while making explicit the previously implicit commitment of taxpayers to support Fannie and Freddie.

The experience during the crisis suggests that additional implicit liabilities would be created by a future housing finance reform that creates an ostensibly private housing finance system. The reason is because future policymakers would doubtless feel obliged to intervene if mortgage financing became difficult to obtain. An ostensibly private housing finance system would thus re-create the implicit government guarantee of the previous failed system.

Conclusion

The use of implicit liabilities is politically attractive in a range of areas. In housing finance, for example, implicit commitments in the past allowed lawmakers to avoid paying for the implicit claims arising from the support that the federal government was expected to provide for Fannie Mae and Freddie Mac if those firms became distressed. That need for taxpayer assistance ultimately materialized in the fall of 2008. Looking forward, policymakers can claim that newly adopted housing finance reform provides for a private system, but, in fact, there is implicit taxpayer exposure that is unpriced.

In general, implicit liabilities have the feature that the full budgetary costs are obscured. This lack of transparency is a political boon for policymakers, who can provide benefits without having to face the attendant costs—at least not yet. Hamilton's paper is thus invaluable in listing and assessing those costs.

Liberty for More: Finance and Educational Opportunities

Ross Levine and Yona Rubinstein

ABSTRACT

U.S. banking reforms—which reduced interest rates—boosted college enrollment rates among able students from middle-class families. We define "able" students as those with learning aptitude scores in the top two-thirds of the U.S. population. We define "middle class" as families in which both parents are not highly educated (more than 12 years of education) and that are neither in the bottom fourth nor in the top 10 percent of the family income distribution in the United States. Our findings suggest that credit conditions, the ability of an individual to benefit from college, and a family's financial and educational circumstances combine to shape college decisions. The functioning of the financial system plays a powerful role in shaping the degree to which a child's educational choices—and hence economic opportunities—are defined by parental income.

Ross Levine holds the Willis H. Booth Chair in Banking and Finance at the Haas School of Business at the University of California, Berkeley. He is also a senior fellow with the Milken Institute and a research associate with the National Bureau of Economic Research. Yona Rubinstein is an Associate Professor in the Managerial Economics and Strategy Group at the London School of Economics.

The authors thank Carlos Avenancio, Maria Carkovic, Raquel Fernández, Erica Field, Bob Hall, Jeff Smith, and seminar participants at the Third Annual Cato Papers on Public Policy Conference for very helpful comments and suggestions.

Liberty for More: Finance and Educational Opportunities

1. INTRODUCTION

Consider two observations about education and income: First, there is a positive association between education and income. On average, people who receive more education go on to earn higher incomes. Second, there is a positive association between parents' income and the education of their children. Thus, higher-income parents tend to have children who receive more education, and that additional education is associated with those children earning higher incomes as adults.[1]

Those two observations motivate some hotly debated questions among social scientists: If the returns to education are so high, why do children from lower-income families obtain less education than children from higher-income families? If education boosts future incomes, what constrains children from lower-income families from attaining more education?

One response to those questions focuses on the costs: lower-income families do not have the money—and borrowing is too expensive—to pay for more education for their children (Becker 1975; Kane 1994; Kane and Rouse 1999; Ellwood and Kane 2000; Belley and Lochner 2007; Brown, Scholz, and Seshadri 2012). From that perspective, financial reforms that lower interest rates will boost education, including college enrollment rates, among children from disadvantaged families that were unable to afford college when credit was more expensive.

A second response stresses the benefits. The returns to higher education for children from disadvantaged families are comparatively low, and so those families invest less in education (Cameron and Heckman 1998, 2001; Shea 2000; Keane and Wolpin 2001;

[1] See Manski and Wise (1983); Behrman and Taubman (1990); Hauser (1993); Kane (1994); Mayer (1997); Cameron and Heckman (1998, 2001); Shea (2000); and Carneiro and Heckman (2002).

Carneiro and Heckman 2002; Keane 2002). That is, if children from lower-income families experience family, community, and grade school environments that are less conducive to their cognitive and noncognitive development than children from higher-income families, then their expected benefits from attending college will be correspondingly lower. Consequently, by the time students are deciding whether to work or go to college, their childhood environments have already determined their expected benefits from going to college. From that perspective, financial reforms that reduce the cost of credit will have a minor effect as disadvantaged families accurately view the expected benefits as low.[2]

To assess those responses empirically, numerous empirical studies use indirect methods (that do not rely on directly observing interest rates) to infer whether interest rates materially shape educational choices. They use indirect methods because of the difficulties associated with identifying exogenous changes in interest rates. As we will review in greater detail, a large body of research devoted to estimating the causal effects of schooling on income (see Card 1999, 2001) has found that the instrumental variable estimates of the returns to schooling exceed the ordinary least squares (OLS) estimates. In theory, credit-constrained individuals will have higher returns to education on the margin than less constrained individuals. That theory implies that if the instrumental variables capture the schooling choices of credit-constrained individuals more than others, the instrumental variable estimates will be larger (Lang 1993; Card 1999, 2001). However, it might be inappropriate to interpret the differences between instrumental variable and OLS estimates as reflecting the effect of the cost of credit on educational choices because other factors can produce those differences (Carneiro and Heckman 2002; Cameron and Taber 2004).

A second but much more limited body of research directly assesses the effect of interest rates on education, but this line of inquiry has serious methodological limitations. In an influential study, Card and Lemieux (2001) find that changes in U.S. interest rates do not account for changes in educational choices over the period

[2] Rather than focusing on reducing the cost of credit, this "benefits" view holds that the most efficacious way to boost college graduation rates among lower-income families is through early childhood interventions that enhance cognitive and noncognitive development and thereby boost the returns to education.

1968–1996. But it is inappropriate to treat the United States as an integrated capital market with a single interest rate during the 20th century with corresponding effects on state-level interest rates. State bank regulations differentially influenced the cost of credit in each state. And states reformed their regulations in different years during the second half of the 20th century. Consequently, both the level and dynamics of interest rates differ across states. Furthermore, since many factors might be correlated with both interest rates and education, it is important to use instrumental variables to identify the effect of interest rates on education.

By integrating labor and financial economics, we contribute to the study of the effect of credit conditions on educational choices in several ways. First, we assess whether state-specific banking reforms that intensified competition among banks and reduced state-specific interest rates increased the probability that students from those states attended college. Previous research on education and credit conditions both failed to recognize that U.S. credit markets were highly segmented because of state-specific regulations on banks for virtually the entire 20th century and failed to exploit the cross-state heterogeneity in the timing of banking reforms that lowered interest rates. In one of the largest—if not the largest—financial regulatory reforms in the history of the United States, every state relaxed geographic restrictions on bank branching—intrastate bank branching reform—during the second half of the 20th century. The state-specific timings of those deregulations were independent of interest rates and education. Although those intrastate bank branch deregulations eased credit conditions, researchers have not—to the best of our knowledge—previously assessed the effect of those regulatory reforms on educational choices.

Second, using state-level bank branch deregulation as an instrumental variable for interest rates, we assess whether this component of state-level interest rates affects the probability that students from that state attend college. However, one must be cautious in drawing sharp inferences from the instrumental variable analyses because bank branch deregulation does not necessarily satisfy both conditions for a valid instrument. Although there is no evidence that educational choices or interest rates influence the timing of bank deregulation in a state, the exclusionary condition might not hold. For example, bank deregulation might accelerate economic

activity and boost the demand for skilled workers, encouraging more students to attend college. Put differently, exogenous bank branch deregulation might reduce interest rates and boost college enrollment rates, but it might not boost college enrollments by lowering interest rates; it might boost college enrollments by increasing the demand for skilled workers. Thus, to interpret the instrumental variable analyses as providing information about the effect of an exogenous change in the cost of borrowing on the decision to attend college, we separately evaluate whether demand-side factors are driving the results.

Third, we assess whether an easing of credit conditions triggered by intrastate bank branch deregulation influenced only those particular children within particular families implied by Becker's (1967) model of human capital accumulation. The model suggests that the effect of lowering interest rates on attending college depends in an interactive manner on family income and the ability of the individual child to benefit from college. That framework suggests that a reduction in interest rates will have a larger effect on high-ability students who would benefit materially from college but whose parents were previously unable to afford college than it will on high-ability children from families that are unconstrained when making decisions about college. Thus, the model predicts that the effect of a change in credit conditions will differ depending on the ability of the individual student to benefit from college and on the ability of the family to pay for college. We assess that prediction using both the simple reduced-form analyses of education and bank deregulation and the instrumental variable analyses that use bank deregulation as an instrument for interest rates.

To make those contributions, we primarily use the National Longitudinal Survey of Youth 1979 (NLSY79) because of its unique characteristics. First, the NLSY79 traces individuals through time so that we know the educational attainment of each person. Second, the NLSY79 contains information on learning aptitude. It gave respondents the Armed Forces Qualification Test (AFQT) in 1980 when they were between the ages of 14 and 22. We use the AFQT score as a measure of learning ability, that is, the ability to benefit from education. Third, the NLSY79 has information about each respondent's family, including family income in 1979 and the educational attainment of both the mother and father. Given the tight connection between

education and income and the problems associated with measuring permanent income using only one year of data, we sometimes use parental education as a proxy measure of the permanent income of the family instead of family income in 1979. Fourth, the NLSY79 has information on two psychometric traits as measured in 1980: (a) self-esteem and (b) the degree to which the person believes that chance, fate, and luck control his or her life. Thus, in assessing the effect of changes in credit conditions on educational choices, we can control for many individual and family characteristics.

To complement the NLSY79, we use the Current Population Survey (CPS), which surveys more people than the NLSY79, but it does not contain information on learning ability, parental education, or personality traits. Thus, we use the CPS to make broader assessments about the effect of bank deregulation on interest rates and the Mincerian[3] returns to education, and we use the NLSY79 to assess how changes in credit conditions influence the decision of individuals to attend college.

We find that intrastate bank deregulation substantially increased the probability that individuals with particular learning abilities and family traits attended college. Specifically, bank deregulation had no effect on students in the lower third of the distribution of learning ability as measured by AFQT; for students for whom the expected benefits of college are low, changes in credit conditions have no appreciable effect on the probability of attending college. But bank deregulation did boost the probability that "able" students—students in the upper two-thirds of the AFQT distribution—go to college. For example, five years after a state deregulated, the probability that able students attend college was 13 percent greater than before deregulation. Moreover, and consistent with theory, an easing of credit conditions has the biggest effect on the able students from families in which both parents have a relatively low level of education (fewer than 12 years of completed schooling). Indeed, for able students from families in which both parents have more than 12 years of education, bank deregulation has no effect. To the extent that parental education is an accurate signal of the family's permanent income (or the family's taste for education), changes in credit conditions have little influence on the decisions of highly educated, affluent parents

[3] See Mincer (1974).

to send their children to college. However, in more disadvantaged families (as measured by parental education), bank deregulation has a large effect: five years after deregulation, able students from disadvantaged families have an almost 20 percent greater probability of attending college. The results are consistent with the view that credit conditions materially influence the educational opportunities of a particular segment of society: able students from disadvantaged families.

When dividing the sample by family income instead of parental education, we find that easing intrastate bank deregulation boosted college enrollment rates among able students from middle-class and upper-middle-class families. Even among students in the upper two-thirds of the AFQT distribution, an easing of credit conditions did not influence children from lower-income families (below the 25th percentile of the income distribution) or high-income families (above the 90th percentile). At those income levels, marginal changes in interest rates did not alter decisions about college. However, for able students from families with incomes between the 50th and 75th percentile of the income distribution, bank deregulation materially altered college decisions.

The results are very similar when we use bank deregulation as an instrumental variable for interest rates. Only for able students from middle-class and upper-middle-class families is the reduction in interest rates associated with an increase in the probability of attending college. For lower-income families or high-income families, such changes in the cost of credit do not influence college decisions. And reductions in interest rates do not increase the probability of attending college among students with AFQT scores in the bottom third of the sample. Consistent with theory, changes in the cost of credit influence a particular but meaningful segment of society.

Finally, we show that our results do not simply reflect the effect of intrastate bank deregulation on the demand for skilled labor. Rather, bank deregulatory reforms boosted college enrollment rates among able students from middle-class and upper-middle-class families, partially by lowering the costs of credit. In particular, a legitimate concern with our analyses is that perhaps branch deregulation boosted economic activity (Jayaratne and Strahan 1996) and thereby boosted the demand for skilled labor. Perhaps that "demand-side"

effect is driving the increase in college enrollments, leading us to misinterpret the findings as arising from a reduction in the "cost of credit." Although reasonable conceptually, the demand-side channel is not the only channel through which bank deregulation increased college enrollments. If the results were purely a demand-side effect, then bank deregulation should boost the demand for college-educated workers and the returns to a college education. But we show that bank deregulation reduced the returns to a college education, which is fully consistent with a reduction in costs boosting the supply of college-educated workers. Although we do not rule out the demand-side channel as a contributing factor, the findings suggest that the supply side mattered too, as deregulation eased credit conditions and that boosted the supply of college-educated workers.

The remainder of this paper is organized as follows: Section 2 sketches the theoretical framework and its empirical predictions. Section 3 provides a literature review and details how we propose to contribute to existing research. Section 4 discusses the data on bank deregulation, interest rates, and education. Section 5 presents the results, and Section 6 concludes.

2. THEORETICAL FRAMEWORK AND EMPIRICAL PREDICTIONS

Building directly on Becker (1967) and Mincer (1974), this section first presents a theoretical model of human capital and then details the empirical predictions emerging from the theory. In its simplest form, the Becker (1967) model assumes that each individual i maximizes the discounted present value of lifetime earnings, $W(s_i)$, by choosing the optimal level of investment in human capital, s_i, which we call "schooling" or "education" but which represents all investments in human capital skills that boost earnings.

$$W(s_i) = \int_{t=s}^{\infty} y(s_i) \exp(-(r_i - \theta_i))\, dt \tag{1}$$

where $y(s_i)$ denotes the annual earnings of an individual with a schooling level s_i; r_i is the interest rate facing individual i, which reflects his or her cost of capital and subjective rate of time preference; and θ_i is the individual's preference for schooling over work. For simplicity, we assume an infinite planning horizon.

To complete the model, let earnings reflect the spot market value of a unit of human capital (P) multiplied by the individual's stock of accumulated human capital, $H(s_i)$:

$$y(s_i) = P \times H(s_i) \tag{2}$$

Further, following Griliches (1977) and more recently Card (2001), we define the human capital production function as follows:

$$H(s_i) = \exp(a_i s_i - \frac{\gamma}{2} s_i^2 + \mu_i \tag{3}$$

where individual i's human capital is positively related to his or her schooling (s_i), ability to benefit from schooling (a_i), and initial level of general skills (μ_i). The beneficial effects of additional schooling face diminishing returns (γ), which we assume to be the same across all individuals for simplicity.

Solving, the optimal level of schooling for individual i (s_i^*) is

$$s_i^* = \frac{a_i - r_i + \theta_i}{\gamma} \tag{4}$$

Across individuals, differences in the optimal amount of education reflect differences in the ability to benefit from education (a_i)—modeled as the technological efficiency with which learning time, effort, and resources augment the value of human capital—the cost of credit (r_i), and the (dis)utility from schooling (θ_i). Clearly, if an individual's marginal benefit from education with respect to future income is relatively large (i.e., a large a_i), then that individual will tend to invest relatively more in schooling than a low-ability person. If an individual's preference for education, θ_i, is relatively high, then such an individual will invest more in education than comparable individuals with weaker tastes for schooling. The model is silent about the source of heterogeneity in the "ability to benefit" from education (Ben-Porath 1967) and the "ability to pay." Separating the ability to benefit from education and the ability to pay for it is challenging, especially since family and community environments affect both.

Although it is appropriate to model human capital as a stock and investment in human capital as a flow, schooling—especially higher-level education—is often a discrete choice. Therefore, there are discrete educational choices, such as attending college, for which the effect of easing credit conditions will depend on the initial conditions facing the family and individual student. For example, a high-ability student in a high-income family that has a strong taste for education

might have an optimal level of schooling that includes college and even a postgraduate degree. In this case, lowering interest rates will not affect the decision to attend college. As another example, a low-ability student from a disadvantaged family with weak tastes for education might have an optimal level of schooling that does not even include graduating from high school. In this case, lowering interest rates is unlikely to affect the student's decision to attend college. However, for some high-ability children from families in which initial interest rates cause the expected costs of college to outweigh the expected benefits, an easing of credit conditions could influence the decision to go to college. Thus, the effect of credit conditions on the decision of an individual student in a particular family to attend college may depend materially on the student's ability to benefit from college, the initial financial conditions facing the family, and its taste for education.

3. THE LITERATURE AND OUR CONTRIBUTION

Why don't lower-income families invest more in the education of their children? As we noted in the Introduction, an enormous literature documents large disparities in high school and college graduation rates across family income groups over the 20th century. Since education is so highly correlated with income, those disparities motivate research on the persistence of inter–income group disparities in education.

We have already noted that the model highlights two major—though not mutually exclusive—explanations for why disadvantaged families invest comparatively little in the education of their children. The first emphasizes the costs: Lower-income families do not have the money to pay for more education and their borrowing costs are high. Those costs hinder lower-income families from providing the same level of education to their kids as higher-income families, perpetuating intergenerational income differences. From this perspective, lowering interest rates will lower the costs of education so that high-ability children from lower-income families can better afford college. Thus, improvements in financial systems can reduce inequalities of opportunity and the inefficient persistence of relative income differences.

The second explanation for why lower-income families do not invest more in education stresses the benefits: The children of

disadvantaged families frequently face lower expected returns to additional education; that is, their a_i's are low. Thus, children from disadvantaged families—which tend to provide family, community, and school environments less conducive to the cognitive and noncognitive development of their children—will disproportionately and accurately view college as a relatively low-return activity. From this perspective, lowering interest rates will not induce lower-income families to invest much more in sending their kids to college.

3.1 Existing Evidence

The evidence on whether the credit conditions influence educational choices is mixed and inconclusive. Given the difficulties associated with measuring the credit conditions facing individuals, a large body of research has used indirect methods—which do not require researchers to observe interest rates or other measures of credit conditions—to draw inferences about the influence of credit conditions on educational choices.

Numerous studies have tackled this question by studying the correlation between educational attainment and family income (or other family characteristics). The positive correlation between educational attainment and family income has been widely interpreted as evidence that borrowing constraints hinder educational choices (see, e.g., Kane 1994; Kane and Rouse 1999; Ellwood and Kane 2000; Belley and Lochner 2007; Brown, Scholz, and Seshadri 2012). However, the step from correlation to causation is a precarious one as family income is also strongly correlated with family resources that foster cognitive and noncognitive traits that boost the ability of a student to benefit from more education. The connection between family resources and the nurturing of cognitive and noncognitive traits that increase the productivity of formal schooling has been emphasized by Cameron and Heckman 1998, 2001; Shea 2000; Heckman and Rubinstein 2001; Keane and Wolpin 2001; Carneiro and Heckman 2002; Keane 2002; Cameron and Taber 2004; and Heckman, Stixrud, and Urzua 2006.

As we have already noted, a large literature finds that the instrumental variable estimates of the return to schooling exceed OLS estimates (see Card 1999, 2001). Credit conditions are one possible source of that difference between the estimates, a point first offered by Becker (1967). In particular, instrumental variable estimates can

be interpreted as estimating the return for those randomly assigned to schooling by the instrument. Finding higher returns using an instrumental variable is consistent with the view that those affected by the "instrument" are credit-constrained (Lang 1993; Card 1999, 2001) and is, therefore, consistent with interest rates curtailing the educational opportunities of lower-income families. Similarly, Shea (2000) finds that family income matters for children's human capital investment in a sample of low-income families, but not for the broader population.[4]

Substantial work, however, challenges the methodological efficacy of these indirect methods for drawing inferences about the effect of credit conditions on educational choices. For example, Cameron and Heckman (1998, 2001) and Carneiro and Heckman (2002) suggest that it is inappropriate to interpret the difference between OLS and instrumental variable estimates in the sample of low-income families as signaling the importance of liquidity constraints, criticizing econometrically the use of invalid instruments and pointing economically to alternative explanations, including sorting for schooling on comparative advantage.

Hence, larger coefficients in instrumental variable regressions of income on education might not imply the existence and effect of interest rates on schooling. That is, without directly measuring exogenous changes in interest rates, it is difficult to distinguish between cross-family differences in interest rates (r_i) and attitudes toward education (θ_i). Furthermore, Cameron and Taber (2004) question the robustness of the instrumental variable results to using alternative instruments. Keane and Wolpin (2001) estimate a structural model of schooling behavior and find that relaxing interest rates tends to increase consumption, not investment in education

A much more limited set of papers assesses the direct linkage between interest rates and schooling decisions. As a leading example, Card and Lemieux (2001) find that changes in U.S. interest rates over the period 1968–1996 do not account for changes in educational choices.

However, the direct approaches taken so far have two key limitations. First, it is inappropriate to treat the United States as

[4] Researchers also examine the effect of targeted credit programs on education, such as the CalGrant program in California for college-bound students (Kane 2003) or Head Start.

an integrated capital market with a single interest rate, especially with regard to household loans during the 20th century. Each state exerted a powerful regulatory role over banks until the mid-1990s so that interest rates and their evolution over time differ markedly by state. Second, it is valuable to identify an exogenous source of variation in credit conditions to assess the effect of interest rates on educational choices. Some third factor, such as aggregate economic activity, could affect both interest rates and education decisions, creating a spurious correlation between them.

3.2 Our Contribution

We propose to contribute to existing research in the following interrelated ways:

First, we will *directly* examine the relationship between educational choices and credit conditions, as measured by exogenous changes in bank regulations that lowered interest rates. This examination contrasts with the large literature that draws inferences about the importance of credit conditions in explaining educational choices through indirect methods, that is, by examining differences between instrumental variable and OLS coefficient estimates of the relationship between wages and education.

Second, we will assess the effect of the exogenous relaxation of regulatory restrictions on bank branching (which lowered interest rates) on college enrollment rates. As we will describe in greater detail, those deregulations occurred across all states in varying years during the second half of the 20th century. Since those state-level regulatory reforms occurred in different years, we control for all national influences by including year fixed effects. Furthermore, whereas past studies take the United States as an integrated financial system with one interest rate, we allow interest rates to differ at the state-year level. This allowance is crucial for drawing accurate inferences about the relationship between credit conditions and educational choices because state regulations heavily and differentially influenced credit conditions across the U.S. states for much of the 20th century, and those regulations were liberalized in different years in different states.

Third, we assess the relationship between interest rates and college enrollment rates, using exogenous cross-state, cross-year variation in bank deregulation as an instrumental variable for interest rates. Unlike much existing work, using instrumental variables is valuable

because interest rates and schooling choices might be simultaneously determined. In these analyses, we argue that bank deregulation is exogenous to educational choices, but we do not claim that bank deregulation influences schooling only through its effect on interest rates; we do not claim that bank deregulation satisfies the exclusionary restriction. In particular, bank deregulation might boost the demand for skilled workers and thereby induce more people to attend college. Despite these limitations, we present evidence that deregulation boosted college enrollment rates by reducing the cost of college, not simply by increasing the demand for skilled workers.

Fourth, we provide an empirical bridge between those researchers who focus on the costs of education and those who focus on the benefits of education in seeking to explain why the children of lower-income families tend to obtain less education. To do so, we will evaluate the effect of easing credit conditions ("costs of education") on an individual's educational choices while differentiating by proxies for the person's learning aptitude ("benefits of education") and the family's initial conditions as measured by family income and the education of the parents. Thus, we will assess how the costs and benefits of college combine to shape an individual's educational choices.[5]

4. DATA: BANK DEREGULATION, INTEREST RATES, AND EDUCATION

Geographic restrictions on banks have their origins in the U.S. Constitution, which limited states from taxing interstate commerce and issuing fiat money. In turn, states raised revenues by chartering banks and taxing their profits. Since states received no charter fees from banks incorporated in other states, state legislatures prohibited the entry of out-of-state banks through interstate bank regulations. To maximize revenues from selling charters, states also effectively granted local monopolies to banks by restricting banks

[5] Furthermore, by differentiating by each individual's cognitive abilities and the traits of each person's parents, the framework advertises an additional line of inquiry: credit conditions can influence the ability of a person to benefit from college. For example, if a reduction in interest rates allows a family to purchase a home in a better school district and that school district enhances the cognitive and noncognitive capabilities of the children, then interest rates can increase the returns from additional education for those children. A reduction in r_i can boost a_i, with a concomitant increase in education and lifetime earnings. We are pursuing this line of research in a companion paper.

from branching within state borders. The intrastate branching restrictions frequently limited banks to operating in one city.

By protecting inefficient banks from competition, geographic restrictions created a powerful constituency for maintaining those regulations even after the original fiscal motivations receded. Indeed, banks protected by those regulations successfully lobbied both the federal government and state governments to prohibit interstate banking and intrastate branching (White 1982).

In the second half of the 20th century, however, technological, legal, and financial innovations diminished the economic and political power of banks benefiting from geographic restrictions. In particular, a series of innovations lowered the costs of using distant banks. Those lower costs reduced the monopoly power of local banks and weakened their ability and desire to lobby for geographic restrictions. For example, the invention of automated teller machines (ATMs), in conjunction with court rulings that ATMs are not bank branches, weakened the geographical link between banks and their clientele. The creation of checkable money market mutual funds made banking by mail and telephone easier, further weakening the power of local bank monopolies. Finally, the increasing sophistication of credit-scoring techniques, improvements in information processing, and the revolution in telecommunications reduced the informational advantages of local bankers, especially with regard to small and new firms.

Those national developments interacted with preexisting state characteristics to shape the timing of bank deregulation across the states and the District of Columbia, as listed in Table 1. As shown by Kroszner and Strahan (1999), deregulation occurred later in states where potential losers from deregulation (small, monopolistic banks) were financially stronger and had a lot of political power. On the other hand, deregulation occurred earlier in states where potential winners of deregulation (small firms) were relatively numerous. Thus, unlike many types of regulatory reforms that occur at the national level, geographic bank deregulation took place on a state-by-state basis over an extended period.

4.1 Interest Rates

To measure the cost of credit, we use data on mortgage rates at the state-year level. Since consumers frequently use their homes as collateral, those rates provide information on general credit

Table 1
Dates of Intrastate Bank Branch Deregulation, by States

State	Date	State	Date
Alabama	1981	Montana	1990
Alaska	1960	Nebraska	1985
Arizona	1960	Nevada	1960
Arkansas	1994	New Hampshire	1987
California	1960	New Jersey	1977
Colorado	1991	New Mexico	1991
Connecticut	1980	New York	1976
District of Columbia	1960	North Carolina	1960
Florida	1988	North Dakota	1987
Georgia	1983	Ohio	1979
Hawaii	1986	Oklahoma	1988
Idaho	1960	Oregon	1985
Illinois	1988	Pennsylvania	1982
Indiana	1989	Rhode Island	1960
Iowa	1999	South Carolina	1960
Kansas	1987	Tennessee	1985
Kentucky	1990	Texas	1988
Louisiana	1988	Utah	1981
Maine	1975	Vermont	1970
Maryland	1960	Virginia	1978
Massachusetts	1984	Washington	1985
Michigan	1987	West Virginia	1987
Minnesota	1993	Wisconsin	1990
Mississippi	1986	Wyoming	1988
Missouri	1990		

Source: Kroszner and Strahan (1999).

conditions. The mortgage interest rate data are based on a monthly survey of major lenders that are asked to report the terms and conditions on all conventional, single-family, fully amortized, purchase-money loans closed in the last five working days of the month. The data exclude Federal Housing Administration–insured and Veterans Administration–guaranteed mortgages, refinancing loans, and balloon loans. The "effective interest rate" includes the amortization of initial fees, points, and charges over a 10-year period, which is the historical assumption of the average life of a mortgage loan and is computed by the Federal Housing Finance Board. We then deflate by the national consumer price index.

4.2 Education and Other Individual-Level Data

The NLSY79 is a nationally representative sample of 12,686 young men and women who were 14–22 years old when they were first surveyed in 1979. Interviewees were initially surveyed annually, and then on a biennial basis after 1994. The NLSY is made up of three subsamples: (a) a random sample of 6,111 noninstitutionalized civilians; (b) a supplemental sample of 5,295 people designed to oversample civilian Hispanics, blacks, and economically disadvantaged whites; and (c) a sample of 1,280 people who were ages 17–21 as of January 1, 1979, and who were enlisted in the military as of September 30, 1978. We use the random sample and the black and Hispanic oversamples and use the weights provided by the NLSY79 to obtain a representative sample of the U.S. population.

In the analyses, we control for information on family background, including family income and the highest grade completed by a person's mother and father. *Family Income in 1979* measures the income of the individual's household in 1979, computed in 2010 dollars. (In the regressions, we divide *Family Income in 1979* by $100,000.) As shown in Table 2, the mean value is about $62,300, and the median value is about $56,400. Some values are missing for *Family Income in 1979*. Consequently, when we use *Family Income in 1979* as a regressor, we impute the sample mean and include a dummy variable that equals 1 for observations in which *Family Income in 1979* is missing. When we use *Family Income in 1979* to divide the sample, we do not impute a value for missing observations and instead use a smaller sample of individuals. Therefore, the number of observations is not identical in these different specifications.

Table 2
Summary Statistics

	Mean	**Median**	**Standard Deviation**
Attended College	0.52	1.00	0.50
Years since Deregulation	3.65	0.00	7.16
AFQT Percentile	51.30	51.88	28.86
External Locus of Control Score	8.48	8.00	2.39
Self-Esteem Score	22.57	22.00	4.07
Mother's Education	11.68	12.00	2.79
Father's Education	11.83	12.00	3.57
Family Income in 1979 (2010 dollars)	$62,302	$56,430	$42,221
Interest Rate	5.29	5.17	2.68

Note: *Attended College* equals 1 if the individual attended college for any period of time and 0 otherwise. *Years since Deregulation* equals the number of years since the state deregulated restrictions on intrastate branch banking and is computed for the year 1979. *AFQT Percentile* is the individual's Armed Forces Qualification Test score as a percentile of the entire National Longitudinal Survey of Youth 1979 (NLSY79) sample, which is measured in 1980, and where 50 is the median of the NLSY79 representative sample. *External Locus of Control Score* (computed in 1980) measures the degree to which individuals believe that external factors such as chance, fate, and luck control their lives relative to the degree that the individual has internal control over his or her life, where values range from 4 to 16—higher values signify more external control; *Self-Esteem Score* (computed in 1980) measures the degree of approval or disapproval of one's self, where values range from 6 to 30— higher values signify greater self-approval. *Mother's* and *Father's Education* measure the number of years of education of the mother and father, respectively. *Family Income in 1979* measures the income of the individual's household in 1979, computed in 2010 dollars. In the regression tables, we divide *Family Income in 1979* by $100,000. *Interest Rate* is the annual real interest rate based on mortgage rates from a monthly survey of major lenders that are asked to report the terms and conditions on all conventional, single-family, fully amortized, purchase-money loans closed in the last five working days of the month. The data exclude FHA-insured and VA-guaranteed mortgages, refinancing loans, and balloon loans. The rate includes the amortization of initial fees, points, and charges over a 10-year period, which is the historical assumption of the average life of a mortgage loan and is computed by the Federal Housing Finance Board. We deflate by the national Consumer Price Index. For variables from the NLSY79, the reported summary statistics use the NLSY79 sample weights.

The major dependent variable is the binary indicator *Attended College*, which equals 1 if the individual attended college and 0 otherwise. This variable equals 1 for individuals who attended college but did not graduate, those who graduated from college, and those who went on to graduate school. As shown in Table 2, about half of the individuals attended college. We focus on whether the person attended college since that focuses on the decision of whether to go to college or work, which is the central decision in the theoretical framework. The results, however, are robust to focusing on whether the individual graduated from college or not.

Key explanatory variables are as follows:

AFQT Percentile is the individual's AFQT score as a percentile of the entire NLSY79 sample, where the *AFQT Percentile* is measured in 1980. A percentile of 50 is the median of the NLSY79 sample, but the median is 51 for our main regression sample because of missing values on parental education. The AFQT is a weighted sum of four tests from the 10-part Armed Services Vocational Aptitude Battery. We use the *AFQT Percentile* as a proxy of an individual's "ability to benefit from additional education." To guarantee that school attendance did not influence AFQT test scores, the AFQT score is standardized by the age of the individual at the time of the test (Cameron and Heckman 1993; Neal and Johnson 1996; and Altonji and Pierret 2001). We exclude observations with missing AFQT scores, parental education, state of residency, and education attainment.

External Locus of Control Score (computed in 1980) measures the degree to which individuals believe that external factors such as chance, fate, and luck control their lives relative to the degree that the individual has internal control over his or her life. Values range from 4 to 16, with higher values signifying perceptions of greater external control. The average value is 8.5.

Self-Esteem Score (computed in 1980) measures the degree of approval or disapproval of one's self. Values range from 6 to 30, with higher values signifying greater self-approval. As shown in Table 2, the mean value of *Self-Esteem Score* is about 22.6, with a standard deviation of 4.

In some specifications, we use the CPS. Specifically, we use the March Annual Demographic Survey files for the sample years 1976–2007, taken from the Integrated Public Use Microdata Series. We use this much larger sample of individuals to compute the Mincerian returns to education as discussed below.

5. RESULTS

We begin by assessing the effect of bank deregulation on interest rates using the following specification:

$$r_{jt} = \alpha D_{jt} + \beta D_{jt}^2 + \delta_j + \delta_t + e_{jt} \tag{5}$$

In the equation, r_{jt} is a measure of real interest rates in state j in year t, δ_j and δ_t are vectors of state and year fixed effects, and e_{jt} is the error term. The variables of interest, D_{jt}, and D_{jt}^2, equal the number of years since state j deregulated (and 0 before state j deregulated) and the square of the number of years since deregulation. In total, we have data for 48 states plus the District of Columbia. Consistent with the literature on bank branch deregulation, we eliminate Delaware and South Dakota because the structure of their banking systems was heavily affected by laws that made them centers for the credit card industry.

Table 3 shows that intrastate bank deregulation was associated with a substantial reduction in interest rates when controlling for year and state effects. The significant negative relationship between interest rates and bank deregulation only emerges when controlling for both year and state effects. This finding is consistent with the view that capital markets were segmented across the states and that one needs to abstract from national fluctuations in credit conditions to identify the independent effect of state-level deregulations on state interest rates.

5.1 Bank Deregulation and College: Reduced Form Results

We next assess the reduced form relationship between the removal of geographic restrictions on banks and college enrollment, where we differentiate individuals by AFQT scores and by parental education or family income. Since family income in one year might be a poor proxy for permanent income and since education is highly correlated with income, parental education might provide a more accurate signal of the family's long-run financial situation.

According to the theoretical framework above, we should consider the linear-in-the-parameters probability model for whether a person attends college (s_{ij}):

$$s_{ij} = \beta_a AFQT_i + \beta_{D1} D_{j,79} + \beta_{D2} D_{j,79}^2 + \beta_X X_{ij} + \varepsilon_{ij} \tag{6}$$

where the subscript i indicates a person and j designates a state. We include one observation per person. We use the AFQT score to proxy

Table 3
Bank Deregulation and Interest Rates

	(1)	(2)	(3)	(4)
Years since Deregulation	−0.007	−0.016	−0.007	−0.025**
	(0.012)	(0.013)	(0.009)	(0.012)
Years since Deregulation Squared	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	5.170***	5.180***	2.020***	2.010***
	(0.068)	(0.091)	(0.052)	(0.071)
Year Effects	No	No	Yes	Yes
State Effects	No	Yes	No	Yes
Observations	1,274	1,274	1,274	1,274
R-Square	0.001	0.009	0.971	0.978

Note: This table presents the results of four regressions, where the dependent variable equals the *Interest Rate*, which is computed at the state-year level. *Interest Rate* is the effective interest on mortgages, which includes amortization of initial fees, points, and changes over a 10-year period and is computed by the Federal Housing Finance Board. Consistent with previous research, the sample includes data on the District of Columbia and all states except Delaware and South Dakota, which are dropped because of large concentrations of credit card banks. The sample covers the period from 1976 through 2002. Regressors: *Years since Deregulation* equals the number of years since the state deregulated restrictions on intrastate branch banking. This number varies at the state-year level. Standard errors are clustered at the state level and corrected using the Huber-White adjustment. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

for an individual's "ability to learn." In many specifications, we split the sample by AFQT score to assess whether the effect of bank deregulation on educational choices differs by a student's "ability to learn." $D_{j,79}$ is the number of years since deregulation for state j in 1979. We choose 1979 because it is the first year of the NLSY79 and because it corresponds to a period in the lives of respondents when interest rates and credit conditions are likely to influence educational choices. Survey respondents are between the ages of 14 and 22 in 1979. By using a quadratic for years since deregulation, we allow for changes in credit conditions to have nonlinear effects on educational choices.

As additional regressors, X_{ij}, we include the following: *External Locus of Control Score* (computed in 1980) measures the degree to which individuals believe that external factors such as chance, fate, and luck control their lives relative to the degree that the individual has internal control over his or her life. *Self-Esteem Score* (computed in 1980) measures the degree of approval or disapproval of one's self. *Mother's Education* and *Father's Education* measure the number of years of education of the mother and father, respectively. *Family Income in 1979* measures the income of the individual's family in 1979, computed in 2010 dollars and divided by $100,000 (as a regressor). In several specifications, we split the sample according to parental education or family income so we can assess whether the effect of deregulation on college enrollment differs by those family characteristics. In all specifications, we control for regional, racial, gender, and year-of-birth effects, and we include a dummy variable that equals 1 if we imputed *Family Income in 1979*.

Tables 4 and 5 provide the regression results, where Table 4 provides the OLS estimates and Table 5 gives the probit results. In both tables, column (1) provides the results for the full sample; columns (2) and (3) provide results splitting the sample between those with *AFQT Percentile* above and below 33, respectively. The NLSY79 sample mean value of *AFQT Percentile* is 50. In columns (4) and (5), we consider only individuals with AFQT scores above 33 and further split the sample by the education of the parents: column (4) includes individuals where either parent has 12 years or fewer of education, and column (5) includes individuals where both parents have more than 12 years of education. Since the dependent variable is binary, we focus on the results using probit regressions. The OLS regressions yield very similar inferences. For the probit analyses, the reported coefficients are the estimated marginal effects, evaluated at the mean values of the regressors. The standard errors are clustered at the state level and corrected for heteroskedasticity using the Huber-White adjustment.

Bank regulation boosted the probability that individuals with particular abilities and family traits attend college. In particular, bank deregulation has no effect on relatively low-ability people, that is, people with AFQT scores lower than 33 (Table 5, column [3]). But deregulation does have a positive effect on "able" students—students with AFQT scores greater than 33 (Table 5, column [2]). For instance,

Table 4

College and Bank Deregulation by Ability and Parents' Education: Ordinary Least Squares

	(1)	(2)	(3)	(4)	(5)
Sample:	**All**	**AFQT >33**	**AFQT <33**	**AFQT >33 and Either Parent ≤12 Years of Education**	**AFQT >33 and Both Parents >12 Years of Education**
Years since Deregulation in 1979	0.025**	0.035***	0.003	0.041***	0.024
	(0.012)	(0.012)	(0.014)	(0.012)	(0.014)
Years since Deregulation in 1979 Squared	−0.001*	−0.002***	0.000	−0.002***	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
AFQT Percentile	0.008***	0.009***	0.009***	0.009***	0.005***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)
External Locus of Control Score	−0.001	−0.003	0.002	−0.004	0.002
	(0.002)	(0.003)	(0.003)	(0.003)	(0.005)
Self-Esteem Score	0.009***	0.007***	0.014***	0.007***	0.008**
	(0.002)	(0.002)	(0.003)	(0.002)	(0.004)
Mother's Education	0.014***	0.021***	0.003	0.017***	0.006
	(0.003)	(0.003)	(0.004)	(0.004)	(0.006)
Father's Education	0.015***	0.018***	0.009**	0.015***	0.014**
	(0.002)	(0.002)	(0.004)	(0.002)	(0.006)
Family Income in 1979	0.153***	0.157***	0.021	0.258***	0.021
	(0.046)	(0.049)	(0.104)	(0.072)	(0.071)
Observations	8,534	4,737	3,797	3,836	901
R-Square	0.334	0.245	0.133	0.198	0.184

Note: This table presents the results of five ordinary least squares regressions, where the dependent variable equals 1 if the individual attended college and 0 otherwise. There is one observation per person. Regressors: *Years since Deregulation in 1979* equals the number of years since the state deregulated restrictions on intrastate branch banking and is computed for the year 1979. *AFQT Percentile* is the individual's Armed Forces Qualification Test score as a percentile of the entire National Longitudinal Survey of Youth 1979 (NLSY79) sample, which is measured in 1980, and where 50 is the median of the NLSY79 representative sample. *External Locus of Control Score* (computed in 1980) measures the degree to which individuals believe that external factors such as chance, fate, and luck control their lives relative to the degree that the individual has internal control over his or her life, where values range from 4 to 16—higher values signify more external control. *Self-Esteem Score* (computed in 1980) measures the degree of approval or disapproval of one's self, where values range from 6 to 30—higher values signify greater self-approval. *Mother's* and *Father's Education* measure the number of years of education of the mother and father, respectively. *Family Income in 1979* measures the income of the individual's household in 1979, computed in 2010 dollars and divided by $100,000. The regression includes regional, racial, and gender fixed effects; the individual's year of birth; and a dummy variable that equals 1 if we imputed *Family Income in 1979* with the sample mean because of missing values. Sample: Besides the core NLSY sample, we also include the supplementary data set on blacks and Hispanics. Standard errors are clustered at the state level and corrected using the Huber-White adjustment. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

Table 5

College and Bank Deregulation by Ability and Parents' Education: Probit

	(1)	(2)	(3)	(4)	(5)
Sample:	**All**	**AFQT >33**	**AFQT <33**	**AFQT >33 and Either Parent ≤12 Years of Education**	**AFQT >33 and Both Parents >12 Years of Education**
Years since Deregulation in 1979	0.031 (0.019)	0.038** (0.017)	0.003 (0.016)	0.047*** (0.014)	0.016 (0.012)
Years since Deregulation in 1979 Squared	−0.002 (0.001)	−0. 002** (0.001)	−0.000 (0.001)	−0.002*** (0.001)	−0.001 (0.001)
AFQT Percentile	0.011*** (0.000)	0.010*** (0.000)	0.010*** (0.001)	0.011*** (0.001)	0.003*** (0.000)
External Locus of Control Score	−0.002 (0.003)	−0.004 (0.003)	0.002 (0.003)	−0.005 (0.004)	0.000 (0.003)
Self-Esteem Score	0.013*** (0.002)	0.009*** (0.002)	0.014*** (0.003)	0.008*** (0.002)	0.005*** (0.002)
Mother's Education	0.021*** (0.005)	0.027*** (0.005)	0.004 (0.005)	0.020*** (0.005)	0.004 (0.005)
Father's Education	0.022*** (0.003)	0.023*** (0.003)	0.010*** (0.004)	0.019*** (0.003)	0.011** (0.005)
Family Income in 1979	0.264*** (0.070)	0.264*** (0.070)	0.019 (0.103)	0.319*** (0.090)	0.042 (0.051)
Observations *R*-Square	8,534 0.29	4,737 0.22	3,797 0.13	3,836 0.16	901 0.29

Note: This table presents the results of five probit regressions, where the dependent variable equals 1 if the individual attended college and 0 otherwise, and where the reported coefficients are the estimated marginal effects, evaluated at the mean values of the independent variables. There is one observation per person. Regressors: *Years since Deregulation in 1979* equals the number of years since the state deregulated restrictions on intrastate branch banking and is computed for the year 1979. *AFQT Percentile* is the individual's Armed Forces Qualification Test score as a percentile of the entire National Longitudinal Study of Youth 1979 (NLSY79) sample, which is measured in 1980, and where 50 is the median of the NLSY79 representative sample. *External Locus of Control Score* (computed in 1980) measures the degree to which individuals believe that external factors such as chance, fate, and luck control their lives relative to the degree that the individual has internal control over his or her life, where values range from 4 to 16—higher values signify more external control. *Self-Esteem Score* (computed in 1980) measures the degree of approval or disapproval of one's self, where values range from 6 to 30—higher values signify greater self-approval. *Mother's* and *Father's Education* measure the number of years of education of the mother and father, respectively; and *Family Income in 1979* measures the income of the individual's household in 1979, computed in 2010 dollars and divided by $100,000. The regression includes regional, racial, and gender fixed effects; the individual's year of birth; and a dummy variable that equals 1 if we imputed *Family Income in 1979* with the sample mean because of missing values. Sample: Besides the NLSY sample, we also include the supplementary data set on blacks and Hispanics. Standard errors are clustered at the state level and corrected using the Huber-White adjustment. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

1 year after deregulation, the probability of attending college rose by about 3.5 percent; 5 years after deregulation, the probability is 13 percent greater than before deregulation; 10 years after deregulation, the probability is 15 percent greater; and 15 years after, the probability is 8 percent greater than it was before the state deregulated restrictions on intrastate branching.

Table 5 shows that the able students from families in which neither parent has more than 12 years of education experience the biggest boost from bank deregulation in the probability of going to college. In particular, bank deregulation has no effect on attending college for individuals with AFQT scores above 33 but who are from families in which both parents have more than 12 years of education. But bank deregulation has a very large effect on attending college for able students from less well-educated parents. The estimated effects are large. For instance, 1 year after deregulation, the probability of attending college rises by about 4.1 percent; 5 years after deregulation, the probability is 16 percent greater than before deregulation; 10 years after deregulation, the probability is 21 percent greater; and 15 years after, the probability is 17 percent greater than it was before the state deregulated restrictions on intrastate branching.

Besides the main results on bank deregulation, the analyses also show that AFQT and parental education are positively associated with higher college enrollment even when splitting the sample by AFQT and parental income. Self-esteem is also a good predictor of whether an individual attends college, even after conditioning on the array of individual, family, regional, and national factors included in the regressions.

Overall, the results from Tables 4 and 5 are consistent with the theoretical model presented above. To the extent that people with AFQT scores in the bottom third of the distribution are unlikely to benefit from college, it is unsurprising that an improvement in credit conditions does not influence their probability of attending college. Similarly, to the extent that able students from well-educated parents have a high probability of attending college regardless of credit conditions, we would not expect bank deregulation to influence their likelihood of attending college either. Rather, theory and the evidence in Tables 4 and 5 indicate that changes in credit conditions influence a particular—though significant—segment of society: able students from more disadvantaged family backgrounds.

Table 6 continues these analyses by splitting the sample by different levels of *Family Income in 1979*, rather than by the education of the parents. In these analyses, we only consider "able" students, that is, students with AFQT scores above 33. We present five probit regressions of different samples, where we consider families with family income of (a) less than the 25th percentile of family income in our sample, (b) more than the 25th percentile, (c) more than the median, (d) more than the 75th percentile, and (e) more than the 90th percentile of family income in our full sample. In this paper, we simply present the probit regressions; the OLS regressions yield similar results.

We find that easing credit conditions—as proxied by intrastate bank deregulation—boosted college enrollments for able students from middle-class and upper-middle-class families. As shown in Table 6, changing credit conditions had no effect on college enrollment for students with AFQT scores above the 33rd percentile but who came from lower-income families (families with incomes below the 25th percentile). And changing credit conditions had no effect on college enrollment for able students from high-income families (families with incomes above the 90th percentile). When we consider people from the middle and upper-middle of the income distribution, bank deregulation exerted a positive effect on college enrollment rates. For families with incomes above the median, the results from equation (3) indicate that 5 years after bank deregulation, an able person's probability of attending college is on average 20 percent greater; 10 years after deregulation, it is 25 percent greater; even 15 years after deregulation, the probability of an able person attending college is 15 percent greater than it was before deregulation.

5.2 Two-Stage Least Squares

We now employ a two-stage least squares (2SLS) probit estimator to examine the effect of interest rates on the probability of attending college. That is, we estimate the following probit equation of whether a person attends college or not (s_{ij}):

$$\text{(7)} \qquad \Pr(s_{ij} = 1 \mid AFQT_i, r_i, X_{ij}) = \Phi(\beta_a AFQT_i + \beta_r r_j + \beta_x X_{ij})$$

where $\Phi(\ )$ is the cumulative distribution function of the unit-normal distribution, r_j is the real interest rate in 1979 in state j, and the other variables (X) are as defined earlier. The excluded instrumental variables for

Table 6

College and Bank Deregulation by Ability and Parents' Income: Probit

	(1)	(2)	(3)	(4)	(5)
Sample:	**AFQT >33 and Family Income <25%**	**AFQT >33 and Family Income >25%**	**AFQT >33 and Family Income >50%**	**AFQT >33 and Family Income >75%**	**AFQT >33 and Family Income >90%**
Years since Deregulation in 1979	0.036	0.043**	0.055***	0.048***	0.018
	(0.048)	(0.019)	(0.014)	(0.006)	(0.014)
Years since Deregulation in 1979 Squared	−.002	−0.002**	−0.003***	−0.002***	−0.001
	(0.003)	(0.001)	(0.001)	(0.000)	(0.001)
AFQT Percentile	0.011***	0.010***	0.009***	0.007***	0.005***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
External Locus of Control Score	−0.016*	−0.005	−0.006	−0.006	−0.007
	(0.009)	(0.004)	(0.004)	(0.006)	(0.005)
Self-Esteem Score	0.006	0.009***	0.007**	0.005	0.005*
	(0.006)	(0.002)	(0.003)	(0.003)	(0.003)
Mother's Education	0.015	0.029***	0.024***	0.023**	0.018***
	(0.011)	(0.005)	(0.008)	(0.009)	(0.006)
Father's Education	0.021***	0.022***	0.020***	0.022***	0.007**
	(0.007)	(0.004)	(0.004)	(0.004)	(0.003)
Family Income in 1979	−1.238	0.365***	0.410***	0.233	0.089
	(1.131)	(0.093)	(0.117)	(0.153)	(0.119)
Observations	867	2,904	1,964	986	409
R-Square	0.23	0.23	0.24	0.27	0.38

Note: This table presents the results of five probit regressions, where the dependent variable equals 1 if the individual attended college and 0 otherwise, and where the reported coefficients are the estimated marginal effects, evaluated at the mean values of the independent variables. There is one observation per person. When designating the sample, *Family Income* > *X%* signifies the sample is restricted to individuals for which their family income levels in 1979 were above the X percentile of the sample. Regressors: *Years since Deregulation in 1979* equals the number of years since the state deregulated restrictions on intrastate branch banking and is computed for the year 1979. *AFQT Percentile* is the individual's Armed Forces Qualification Test score as a percentile of the entire National Longitudinal Study of Youth 1979 (NLSY79) sample, which is measured in 1980, and where 50 is the median of the NLSY79 representative sample. *External Locus of Control Score* (computed in 1980) measures the degree to which individuals believe that external factors such as chance, fate, and luck control their lives relative to the degree that the individual has internal control over his or her life, where values range from 4 to 16—higher values signify more external control. *Self-Esteem Score* (computed in 1980) measures the degree of approval or disapproval of one's self, where values range from 6 to 30—higher values signify greater self-approval. *Mother's* and *Father's Education* measure the number of years of education of the mother and father, respectively. *Family Income in 1979* measures the income of the individual's household in 1979, computed in 2010 dollars and divided by $100,000. The regression includes regional, racial, and gender fixed effects; the individual's year of birth; and a dummy variable that equals 1 if we imputed *Family Income in 1979* with the sample mean because of missing values. Sample: Besides the core NLSY sample, we also include the supplementary data set on blacks and Hispanics. Standard errors are clustered at the state level and corrected using the Huber-White adjustment. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

r_j are $D_{j,79}$ and $D^2_{j,79}$. Recall from Table 3 that bank deregulation helps account for cross-state, cross-time variation in interest rates. The first stage here is different from Table 3 because there is no time variation; there is one observation per person. However, the coefficient estimates from the first stage correspond with those presented in Table 3. In presenting the 2SLS probit estimates, Table 7 uses a structure similar to Table 6, that is, we split the sample by an individual's AFQT scores and family income.

For able students from middle-class and upper-middle-class families, we find that a reduction in interest rates from bank deregulation during an individual's formative years is associated with an increase in the probability of attending college. For the full sample of individuals, there is no significant relationship between attending college and interest rates (Table 7, column [1]). There is no significant relationship between interest rates and attending college when we only examine able students from lower-income families (families with incomes below the 25th percentile) or when we only examine able students from high-income families (families with incomes above the 90th percentile). Only when we consider able students from middle-class or upper-middle-class families—that is, when we restrict the sample to able students from families with incomes above the median or above the 75th percentile—do we find that changes in interest rates triggered by bank deregulation are negatively and significantly associated with college attendance.

The economic magnitude of the relationship between interest rates and college attendance is large for the subsample of students with AFQT scores above the 33rd percentile who are from middle-class and upper-middle-class families. The coefficient estimate for the sample of able students from families with incomes above the median indicates that a reduction in real interest rates of 1 percentage point is associated with an increase of almost 40 percent in the probability of attending college. For a not-inconsequential segment of society, credit conditions are importantly linked with college attendance.

As noted earlier, we must interpret the 2SLS estimates cautiously because the instruments are unlikely to satisfy the exclusion restriction. Although there is no evidence that educational choices influence the timing of intrastate branch deregulation, there is evidence that intrastate branch deregulation influenced many features of

Table 7

College and Interest Rates by Ability and Parents' Income: Instrumental Variable Probit

	(1)	(2)	(3)	(4)	(5)	(6)
Sample:	**All**	**AFQT >33 and Family Income <25%**	**AFQT >33 and Family Income >25%**	**AFQT >33 and Family Income >50%**	**AFQT >33 and Family Income >75%**	**AFQT >33 and Family Income >90%**
Interest Rate	−0.184	−0.049	−0.258	−0.385**	−0.465***	−0.151
	(0.153)	(0.384)	(0.171)	(0.150)	(0.079)	(0.275)
AFQT Percentile	0.027***	0.030***	0.028***	0.030***	0.030***	0.038***
	(0.001)	(0.003)	(0.002)	(0.002)	(0.002)	(0.005)
External Locus of Control Score	−0.007	−0.043*	−0.015	−0.021	−0.029	−0.062
	(0.007)	(0.025)	(0.010)	(0.013)	(0.025)	(0.048)
Self-Esteem Score	0.032***	0.016	0.026***	0.024**	0.025	0.047*
	(0.005)	(0.017)	(0.007)	(0.012)	(0.016)	(0.026)
Mother's Education	0.050***	0.043	0.085***	0.078***	0.097**	0.146**
	(0.014)	(0.029)	(0.015)	(0.024)	(0.040)	(0.064)
Father's Education	0.054***	0.055***	0.065***	0.068***	0.100***	0.064**
	(0.007)	(0.020)	(0.011)	(0.014)	(0.016)	(0.029)

(continued)

Table 7
(continued)

	(1)	(2)	(3)	(4)	(5)	(6)
Sample:	**All**	**AFQT >33 and Family Income <25%**	**AFQT >33 and Family Income >25%**	**AFQT >33 and Family Income >50%**	**AFQT >33 and Family Income >75%**	**AFQT >33 and Family Income >90%**
Family Income in 1979	0.674***	−3.341	1.036***	1.309***	0.998	0.609
	(0.180)	(3.060)	(0.269)	(0.392)	(0.668)	(0.908)
Observations	6,909	867	2,904	1,964	986	409

Note: This table presents the results of six probit regressions, where the dependent variable equals 1 if the individual attended college and 0 otherwise, and where the reported coefficients are the estimated marginal effects, evaluated at the mean values of the independent variables. There is one observation per person. When designating the sample, *Family Income* > X% signifies the sample is restricted to individuals for which their family income levels in 1979 were above the X percentile of the sample. Regressors: *Years since Deregulation in 1979* equals the number of years since the state in which the individual was living in 1979 deregulated restrictions on intrastate branch banking and is computed for the year 1979. *AFQT Percentile* is the individual's Armed Forces Qualification Test score as a percentile of the entire National Longitudinal Survey of Youth 1979 (NLSY79) sample, which is measured in 1980, and where 50 is the median of the NLSY79 representative sample. *External Locus of Control Score* (computed in 1980) measures the degree to which individuals believe that external factors such as chance, fate, and luck control their lives relative to the degree that the individual has internal control over his or her life, where values range from 4 to 16—higher values signify more external control; *Self-Esteem Score* (computed in 1980) measures the degree of approval or disapproval of one's self, where values range from 6 to 30—higher values signify greater self-approval. *Mother's* and *Father's Education* measure the number of years of education of the mother and father, respectively. *Family Income in 1979* measures the income of the individual's household in 1979, computed in 2010 dollars and divided by $100,000. The regression includes regional, racial, and gender fixed effects; the individual's year of birth; and a dummy variable that equals 1 if we imputed *Family Income in 1979* with the sample mean because of missing values. Sample: Besides the core NLSY sample, we also include the supplementary data set on blacks and Hispanics. Standard errors are clustered at the state level and corrected using the Huber-White adjustment. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

the economy. Branch deregulation accelerated economic activity (Jayaratne and Strahan 1996), reduced income inequality by disproportionately helping the poor (Beck, Levine, and Levkov 2010), and reduced discrimination against black workers (Levine, Levkov, and Rubinstein 2014). Thus, bank deregulation might influence college enrollment rates through a variety of channels beyond its effect on interest rates. It could very well be that deregulation boosted the demand for skilled workers and that it is this "demand-side" effect that drives the increase in college enrollment, not the "cost-side" effect associated with the drop in interest rates.

Although the exclusionary restriction might not hold, two observations suggest that bank deregulation is pushing up college enrollment rates by reducing the cost of college, not just by increasing the demand for skilled workers. First, Beck, Levine, and Levkov (2010) show that bank deregulation increased the demand for unskilled workers, not skilled workers. Thus, bank deregulation does not seem to have increased the demand for skilled workers, which is the starting point of the demand-side story.

Second, we now assess whether bank deregulation increased or decreased the returns to education. According to the demand-side story, bank deregulation should increase the demand for skilled labor and hence the returns to education. In contrast, the supply-side view suggests that bank deregulation lowered the costs of a college education, boosted the supply of college-educated workers, and thereby lowered the returns to education. We assess which prediction holds in the data.

In Table 8, we regress the Mincerian returns to education on bank deregulation over the period 1976–2002 using the following specification:

$$MR_{jt} = \alpha D_{jt} + \beta D_{jt}^2 + \delta_j + \delta_t + e_{jt} \tag{8}$$

The dependent variable, MR_{jt}, equals the Mincerian returns to education in state j during year t and is computed by regressing—by state and year—log wages on years of education and a quartic in potential work experience and by collecting the estimated coefficient on years of education. To compute MR_{jt}, we use the CPS March Supplements and make the computations over the years 1976–2002 for the sample of full-time, full-year, white males between the ages of 25 and 55 while we exclude people living in group quarters and working in either the military or agriculture. We use

Table 8
Bank Deregulation and the Returns to Education

	(1)	(2)	(3)	(4)
Years since Deregulation	0.0021***	0.0035***	0.0004	−0.0010**
	(0.0006)	(0.0003)	(0.0005)	(0.0004)
Years since Deregulation Squared	0.0000**	0.0000**	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Constant	0.0733***	0.0618***	0.0592***	0.0657***
	(0.0027)	(0.0016)	(0.0032)	(0.0025)
Year Effects	No	No	Yes	Yes
State Effects	No	Yes	No	Yes
Observations	1,274	1,274	1,274	1,274
R-Square	0.134	0.522	0.370	0.602

Note: This table presents the results of four regressions, where the dependent variable is *Returns to Education*. For each state-year, *Returns to Education* is computed by regressing log wages on years of education and a quartic in potential work experience and by collecting the estimated coefficient on years of education, as we use data from the CPS March Supplement, over the years 1976–2002, for the sample of full-time, full-year, white males between the ages of 25 and 55, and as we exclude people living in group quarters and working in either the military or agriculture. Consistent with previous research, the sample includes data on the District of Columbia and all states except Delaware and South Dakota, which are dropped because of large concentrations of credit card banks. Regressors: *Years since Deregulation* equals the number of years since the state deregulated restrictions on intrastate branch banking and is computed at the state-year level. Standard errors are clustered at the state level and corrected using the Huber-White adjustment. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

the CPS, rather than the NLSY79, because the CPS samples a much larger cross section of individuals, and we do not need the longitudinal nature of the NLSY79 to compute the returns to education at the state-year level.

As shown in Table 8, bank deregulation reduced the returns to education, after controlling for state and year effects. That reduction is consistent with the cost-side channel playing an important role, whereby bank deregulation lowered the costs of education, shifted out the supply curve of skilled workers, and reduced the returns to education. The findings are inconsistent with a purely demand-side

story in which bank deregulation only boosted the demand for skilled labor, shifted out the demand curve, and increased the returns to education.

The results presented in Tables 7 and 8 suggest that by lowering the costs of a college education, bank deregulation boosted college enrollment rates among able students from middle-class and upper-middle-class families. Although those results do not indicate that bank deregulation increased college enrollment *only* by lowering costs, the results do suggest that bank deregulation increased college enrollments by lowering costs. Although we focus on the effect of deregulation on interest rates, it is possible that deregulation boosted college enrollment rates by increasing family incomes and thereby reducing the costs of funding a child's college education. Although the income channel is feasible, Beck, Levine, and Levkov (2010) show that deregulation boosted the incomes of families in the lower third of the income distribution, and we find that the major effect of deregulation on college enrollment rates occurs in families between the median and 90th percentiles of the income distribution. Thus, even though we cannot nail down the interest rate channel per se, the evidence indicates that improvements in credit conditions triggered by bank deregulation increased college enrollment rates.

6. CONCLUSIONS

In this paper, we find that intrastate bank branch deregulation, which lowered interest rates, boosted college enrollments among able students from middle-class and upper-middle-class families. Our findings suggest that credit conditions, the ability of an individual to benefit from college, and a family's financial circumstances combine to shape decisions about attending college. Banking reforms that ease credit conditions boost college enrollment rates among a significant portion of the population.

The analyses suggest that the functioning of the financial system exerts a powerful influence on shaping economic opportunities, as emphasized by Levine (2005). Although many factors shape the economic opportunities available to a child, affording a good education is one of them. The results presented in this paper indicate that improvements in the functioning of the financial system help make education affordable to more students.

REFERENCES

Altonji, Joseph G., and Charles R. Pierret. 2001. "Employer Learning and Statistical Discrimination." *Quarterly Journal of Economics* 116 (1): 313–50.

Beck, Thorsten, Ross Levine, and Alexey Levkov. 2010. "Big Bad Banks? The Impact of U.S. Branch Deregulation on Income Distribution." *Journal of Finance* 65 (5): 1637–67.

Becker, Gary S. 1967. *Human Capital and the Personal Distribution of Income*. Ann Arbor: University of Michigan Press.

———. 1975. *Human Capital*. 2nd ed. New York: Columbia University Press.

Behrman, Jere R., and Paul Taubman. 1990. "The Intergenerational Correlation between Children's Adult Earnings and Their Parents' Income: Results from the Michigan Panel Survey of Income Dynamics." *Review of Income and Wealth* 36 (2): 115–27.

Belley, Philippe, and Lance Lochner. 2007. "The Changing Role of Family Income and Ability in Determining Educational Achievement." Working paper, Kansas State University, Manhattan.

Ben-Porath, Yoram. 1967. "The Production of Human Capital and the Life Cycle of Earnings." *Journal of Political Economy* 75 (4): 352–65.

Brown, Meta, John K. Scholz, and Ananth Seshadri. 2012. "A New Test of Borrowing Constraints for Education." *Review of Economics Studies* 79 (2): 511–38.

Cameron, Stephen V., and James J. Heckman. 1993. "The Nonequivalence of High School Equivalents." *Journal of Labor Economics* 11 (1): 1–47.

———. 1998. "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males." *Journal of Political Economy* 106 (2): 262–333.

———. 2001. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males." *Journal of Political Economy* 109 (3): 455–99.

Cameron, Stephen V., and Christopher Taber. 2004. "Estimation of Educational Borrowing Constraints Using Returns to Schooling." *Journal of Political Economy* 112 (1): 132–82.

Card, David. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, vol. 3A, ed. Orley Ashenfelter and David Card, pp. 1801–63. Amsterdam: Elsevier Science, North-Holland.

———. 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica* 69 (5): 1127–60.

Card, David, and Thomas Lemieux. 2001. "Dropout and Enrollment Trends in the Post-War Period: What Went Wrong in the 1970s?" In *Risky Behavior among Youths: An Economic Analysis*, ed. Jonathan Gruber, pp. 439–82. Chicago: University of Chicago Press.

Carneiro, Pedro, and James J. Heckman. 2002. "The Evidence on Interest Rates in Post-Secondary Schooling." *Economic Journal* 112: 989–1018.

Ellwood, David T., and Thomas J. Kane. 2000. "Who Is Getting a College Education? Family Background and the Growing Gaps in Enrollment." In *Securing the Future: Investing in Children from Birth to College*, ed. Sheldon Danziger and Jane Waldfogel, pp. 283–324. New York: Russell Sage Foundation.

Griliches, Zvi. 1977. "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica* 45 (1): 1–22.

Hauser, Robert M. 1993. "Trends in College Attendance among Blacks, Whites and Hispanics." In *Studies of Supply and Demand in Higher Education*, ed. Charles T. Clotfelter and Michael Rothschild, pp. 61–120. Chicago: University of Chicago Press.

Heckman, James J., and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review* 91 (2): 145–49.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411–82.

Jayaratne, Jith, and Philip E. Strahan. 1996. "The Finance-Growth Nexus: Evidence from Bank Branch Deregulation." *Quarterly Journal of Economics* 111 (3): 639–70.

Kane, Thomas J. 1994. "College Entry by Blacks since 1970: The Role of College Costs, Family Background, and the Returns to Education." *Journal of Political Economy* 102 (5): 878–911.

———. 2003. "A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going." NBER Working Paper no. 9703, National Bureau of Economic Research, Cambridge, MA.

Kane, Thomas J., and Cecilia Elena Rouse. 1999. "The Community College: Educating Students at the Margin between College and Work." *Journal of Economic Perspectives* 13 (1): 63–84.

Keane, Michael P. 2002. "Financial Aid, Borrowing Constraints, and College Attendance: Evidence from Structural Estimates." *American Economic Review* 92 (2): 293–97.

Keane, Michael P., and Kenneth I. Wolpin. 2001. "The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment." *International Economic Review* 42 (4): 1051–1103.

Kroszner, Randall, and Philip E. Strahan. 1999. "What Drives Deregulation? Economics and Politics of the Relaxation of Bank Branching Deregulation." *Quarterly Journal of Economics* 114 (4): 1437–67.

Lang, Kevin. 1993. "Ability Bias, Discount Rate Bias, and the Return to Education." Working paper, Boston University.

Levine, Ross. 2005. "Finance and Growth: Theory and Evidence." In *Handbook of Economic Growth*, ed. Philippe Aghion and Steven Durlauf, pp. 865–934. Amsterdam: North-Holland Elsevier Publishers.

Levine, Ross, Alexey Levkov, and Yona Rubinstein. 2014. "Bank Deregulation and Racial Inequality in America." *Critical Finance Review* 3 (1): 1–48.

Manski, Charles F., and David A. Wise. 1983. *College Choice in America*. Cambridge, MA: Harvard University Press.

Mayer Susan E. 1997. *What Money Can't Buy*. Cambridge, MA: Harvard University Press.

Mincer, Jacob. 1974. *Schooling, Experience and Earnings*. New York: Columbia University Press.

Neal, Derek A., and William R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy* 104 (5): 869–95.

Shea, John. 2000. "Does Parent's Money Matter?" *Journal of Public Economics* 77 (2): 155–84.

White, Eugene N. 1982. "The Political Economy of Banking Regulation, 1864–1933." *Journal of Economic History* 42 (1): 33–40.

Comment

Raquel Fernández

Starting with the well-known observation that family income and education are positively correlated, this paper asks an important question: to what extent does that correlation reflect the effect of credit constraints? In particular, the paper seeks to examine the link between college attendance and credit constraints. The obstacle to answering that question lies in the difficulty of reliably identifying the credit channel.

The authors' contribution is the use of the variation in the timing of intrastate bank deregulation as a way to obtain "exogenous" variation in the competitiveness of the banking system and hence in the interest rates it charges. This route is potentially promising, and the authors do a good job of dealing with various possible objections. Nonetheless, a few potentially important problems lie in interpreting their results. I discuss those below.

As a first step, the authors show that, after controlling for year and state fixed effects, the (mortgage) interest rate is a negative function of the number of years a state has been deregulated (Table 3).[1] This result is necessary in order for the analysis that follows to make sense—ceteris paribus, states had lower interest rates after deregulation. I would have liked to have seen plots of the data using time windows of varying length to have a feel for the variation and size of the effect. In general, the entire paper would have been improved by presenting graphs showing the variation in the data used to support the main results.

The authors do not justify their choice of the mortgage interest rate as the relevant rate facing individuals. What is the marginal interest rate for loans for most individuals regarding college decisions?

Raquel Fernández is professor of economics at New York University.

[1] Although they include a quadratic specification in *Years since Deregulation*, the estimated relationship turns out to be linear.

To the extent that such loans are mainly federal student loans, the rate will not vary by state. If, on the other hand, the ability of individuals to prosper from college relies on a good primary and secondary education, then where one lives is the main determinant of education quality at this level. In that case, the mortgage interest may be the best indicator of how expensive families find it to live in communities with good schools. I would urge the authors to explore that possibility in future research.

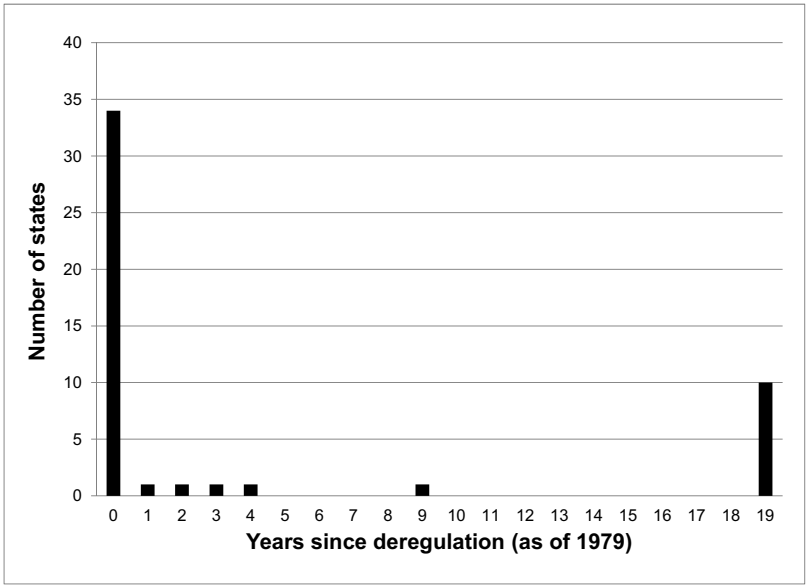
The main analysis consists of ordinary least squares and probit regressions in which the dependent variable is whether an individual *i* from state *j* attended college. The data are from the National Longitudinal Survey of Youth 1979, which follows a sample of individuals who were between the ages of 14 and 22 in 1979. The regressions control for year of birth, gender, parental income in 1979 (or parental education), Armed Forces Qualification Test score (age normalized), and some individual attitudes (self-esteem and locus of control).[2] The main variable is the number of years since deregulation for state *j* (henceforth $d_{j,79}$) as of 1979 if deregulation occurred; otherwise, it takes the value of 0.

Although in principle the variable $d_{j,79}$ is sensible, a quick look at the years in which deregulation occurred (provided in the paper's Table 1) raises serious doubts. A large group of states (10) deregulated in 1960 (hence they would be coded 19), and an even larger set of states (over 30) deregulated in 1979 or after (hence they would be coded 0). As shown in my Figure 1, this variable leaves very little variation across states to "explain" the variation in college outcomes. There appears to be no good reason to code the variable in this way.

A better approach would be to conduct a difference-in-difference analysis by examining the difference in college outcomes for the cohort of individuals too old (say, age 22 or older) to be affected by the state's deregulation in a given year relative to the potentially affected younger cohort in that same state and compare it with the equivalent cohort difference in college attendance in a state that did not deregulate. Unfortunately, the National Longitudinal Survey of Youth 1979 has data for individuals born in a narrow age range. Nonetheless, given that a sizable number of states deregulated between 1975 and 1985 (15), such an analysis may well be feasible.

[2] Why these attitudes should be included in the controls is unclear, and I would have preferred to have seen the regressions without them.

Figure 1
Distribution of $D_{j,79}$



Quantitatively, the results are rather suspect, raising further doubts about the main dependent variable. As indicated previously, Table 3 (column [4]) regressed the mortgage interest rate on the number of years the state had deregulated. The coefficient on this variable indicates that after 10 years the real interest rate is expected, on average, to decrease by 0.2. Given that the real interest rate has a mean of 5.2, it is difficult to believe that college decisions are so sensitive to those small changes. In fact, Table 5 (which contains the main probit results) indicates that one would expect a 21 percent increase in the probability of attending college by high-ability individuals from less educated (hence presumably lower-income) families. That is a very large response to a small change in interest rates!

One objection I raised to an earlier draft of this paper was the possibility that the demand side could be playing a large role. That is, intrastate bank deregulation has been previously shown to have been associated with more startups and increased entry and exit of

firms (Kerr and Nanda 2009); more entrepreneurs (Black and Strahan 2002); lower barriers to entry, especially in bank-dependent industries (Cetorelli and Strahan 2006); and a smaller black–white wage gap (Levine, Levkov, and Rubinstein 2011). If bank deregulation increased the demand for skilled workers, it would also affect the attractiveness of college. The authors now address that concern by showing (Table 8, column [4]) that the return to years of schooling (for full-time-employed white men) is negative once (and only once) they include year and state fixed effects. It would have been interesting to see whether those results could have been replicated using an indicator for some college and above. In any case, as the authors acknowledge, this analysis cannot eliminate the possibility that demand played an important role but that it was overridden by the supply response, resulting in lower wages for college-educated workers.

Next, I turn to some comments regarding the analytical framework for the results. It is interesting to note that the authors never define what it means to be credit-constrained. Is it an inability to borrow on the part of lower-income individuals? Is it their facing higher interest rates than others face? Note that a decrease in interest rates should, ceteris paribus, make investing in human capital more attractive for everyone. Hence, simply observing a positive response does not allow one to conclude that individuals are credit-constrained. For the purpose of their paper—especially given their finding that deregulation increases the probability of attending college for individuals in the upper two-thirds of the Armed Forces Qualification Test distribution with at least one parent with fewer than 12 years of education—I think that a model that distinguishes between the interest rate on borrowing versus the one on saving and that models college as a discrete choice from the outset may be more useful. Below I sketch a very simple model that delivers results consistent with the authors' empirical results.

Consider the college decision of an individual with parental income y and endowed with ability a. Suppose that by going to college the individual earns ay_H, and by not going to college he or she earns y_L. The cost of college is the same for everyone: $c > 0$. Next, assume that individuals can borrow at rate r_b and save at rate r_s, where $r_b > r_s$. Lastly, assume for simplicity that individuals maximize household consumption. That maximized household consumption takes place in the second period, after the individual has attended college. It is simply the sum of earnings plus savings minus debt repayment.

A first observation is that $r_b > r_s$ implies that individuals will always finance college from parental income rather than by borrowing, to the extent possible. It follows that for individuals with parental income high enough ($y \geq c$), the individual will attend college if and only if $ay_H - y_L \geq c\,(1 + r_s)$.[3] For individuals with lower incomes ($y < c$), those individuals will attend college if and only if $ay_H - y_L \geq c\,(1 + r_b) - y\,(r_b - r_s)$.

It follows then that the model predicts:

- Changes in r_b do not affect their college attendance.
- Family income does not play a role in college attendance.
- Only individuals of sufficiently high ability attend college.

Those results are in accordance with what the authors find (see their Tables 4–7). For individuals from lower-income families, on the other hand, the model predicts the following:

- Decreases in r_b will increase college attendance.
- Higher family income increases college attendance.
- A higher minimum ability is required for such students to attend college than for students from wealthy families.

Those predictions are also in line with the empirical results in Levine and Rubinstein's Tables 4–7.

Let me conclude by stating that this paper raises interesting and provocative questions about the ways in which access to financing impinges on education decisions. I hope that the authors continue to work on this important topic.

REFERENCES

Black, Sandra E., and Philip E. Strahan. 2002. "Entrepreneurship and Bank Credit Availability." *Journal of Finance* 57 (6): 2807–33.

Cetorelli, Nicola, and Philip E. Strahan. 2006. "Finance as a Barrier to Entry: Bank Competition and Industry Structure in Local U.S. Markets." *Journal of Finance* 61 (1): 437–61.

Kerr, William R., and Ramana Nanda. 2009. "Democratizing Entry: Banking Deregulations, Financing Constraints, and Entrepreneurship." *Journal of Financial Economics* 94 (1): 124–49.

Levine, Ross, Alexey Levkov, and Yona Rubinstein. 2011. "Racial Discrimination and Competition." CEP Discussion Paper no. 1069. Centre for Economic Performance, London School of Economics and Political Science, London.

[3] An individual with family income y and ability a who attends college consumes $(y - c)\,(1 + r_s) + ay_H$, whereas consumption from not attending college is $y\,(1 + r_s) + y_L$. The inequality follows immediately.

Comment

Erica Field

In their paper, Ross Levine and Yona Rubinstein explore an extremely important question in the literature on the economics of education: what are the major barriers to college attendance in developed countries? Despite high returns to higher education in the United States, rates of college enrollment remain low among many segments of the population and were low in absolute terms throughout much of history. Although those low rates have many possible explanations, the bulk of them can be classified into either cost constraints (affordability combined with credit constraints) or low returns (essentially poor preparation of those educated in lower-income school districts).

Here, the authors attempt to isolate the role of cost constraints by using the natural experiment provided by interstate banking deregulation, which occurred between 1960 and 1999, and was associated with a sharp reduction in interest rates on consumer lending that resulted from greater competition in the banking sector. Although this instrument has been used numerous times in the past to look at the effect of interest rates on local economic activity, Levine and Rubinstein are the first to use it to isolate the causal effect of lending rates on schooling decisions. The results are also novel insofar as they link large-scale financial reforms to changes in schooling opportunity across income groups, and they demonstrate clear heterogeneity with regard to impact of reforms across household types (low-income) and individuals (high-ability).

The fact that the natural experiment is not new in the literature in no way detracts from the importance of the contribution of looking at its effect on human capital. Sources of arguably exogenous variation

Erica Field is an associate professor of economics at Duke University and a faculty research fellow with the National Bureau of Economic Research.

in factors influencing schooling choices are hard to come by, and they should be taken advantage of whenever appropriate to delve deeper into the topic. However, it does imply that we already have hard evidence from existing publications that the timing of interstate banking deregulation does not work well as a means of isolating changes in the cost of schooling on account of the myriad other economic effects of deregulation that have been documented in the literature. In particular, the banking reforms have been associated with increases in rates of economic growth (Jayaratne and Strahan 1996; Huang 2008), higher rates of entrepreneurship (Black and Strahan 2002; Kerr and Nanda 2009), and increases in economic volatility (Morgan, Rime, and Strahan 2004; Demyanyk, Ostergaard, and Sørensen 2007), among other effects. As such, there is no way to be sure that the effects documented in Levine and Rubinstein's paper operate exclusively through changes in the cost of borrowing to pay for college.

The authors are well aware of that fact and do not push the instrumental variables specification particularly hard, but this feature of the identification strategy nonetheless muddies our interpretation of the associations they uncover between banking deregulation and college attendance and calls into question some of the paper's claims. If one assumes that the timing of deregulation is indeed independent of other time trends in economic activity, there is still a multitude of ways in which interest rates could indirectly give rise to changes in schooling choices beyond their potential direct price effects. Are parents' wages (immediate or anticipated future wage profiles) increasing because of greater local economic activity? Is the local economic stimulus associated with an increase in public spending on education or with a convergence in economic opportunities that generate convergence in schooling outcomes (which could show up as an increase in enrollment rates among those outside of the top decile)? Or does an improvement in the local economy motivate kids to enroll in college or parents to invest in kids' college education? Yet another category of explanation unaccounted for in the paper is taste for higher education, which varies with family background and can change with local economic conditions through changes in social norms or peer effects. This channel would boost the effect of any reform that lowered direct cost constraints, for instance.

The possibility of multiple channels of influence is also potentially relevant for our interpretation of the heterogeneity result.

For instance, might it be the case that lower interest rates made it possible for *wealthier* kids to attend private or out-of-state colleges, thereby decreasing competition for spots at lower-cost schools, which led to higher enrollment among the middle class? Although still a relaxation of the cost constraint, if that were the case, it would be incorrect to attribute changes in the behavior of middle-class students to a direct response to relaxed credit constraints.

Given all of those possibilities, two-stage least squares (2SLS) estimates are impossible to interpret, and a reduced form association between deregulation and college enrollment is the only specification that makes sense to use in the paper. It is also the case that the 2SLS estimates are implausibly large, which is hardly surprising given the many potential avenues of influence that are being attributed to interest rate changes in this specification. In this instance, the magnitudes of the estimates should offer guidance as to the plausibility of the assumptions required of the 2SLS specification, and here the numbers make it seem particularly unlikely that the documented effects operate only through the cost of borrowing for college.

The authors certainly acknowledge the fact that many possible channels of influence exist and cite the precedent papers that make the indirect mechanisms impossible to ignore. However, they also argue in favor of the role of credit constraints and college affordability by presenting evidence against alternative pathways. Unfortunately, the evidence is not comprehensive enough to rule out all competing channels, and it even increases the plausibility of certain channels. For instance, they argue that evidence from previous work shows that deregulation was associated with an increase in the demand for unskilled workers, and therefore it is unlikely that enrollment rates are increasing in response to an increase in returns to college. Although this argument makes sense, the existing evidence on demand for unskilled workers makes it even more likely that the effects found by Levine and Rubinstein reflect an increase in the wages of low-skilled parents. Alternatively, an increase in the availability of low-skilled jobs may have made it easier for students to work their way through college.

If one is to gauge the potential for direct responses to the availability of credit, direct evidence would be useful here too. Were enough students from middle-income families actually taking out loans to finance higher education in 1980 for this explanation to be plausible?

As researchers gather direct evidence that college affordability has increased, focusing on the extensive margin would be particularly useful. That is, it is hard to believe that households are very sensitive to small changes in the interest rate on borrowing. A more intuitive story is that the response was driven by changes in the availability of college loans, so it would be particularly helpful to show that availability changed significantly for the group that exhibits the largest change in behavior.

Because the responses documented in the paper are highly lagged, it is particularly difficult to rule out indirect effects of interest rates on schooling decisions that operate through changes in local economic activity. This particular interpretation would gain credence by showing that the response pattern in fact exhibits a discrete jump in schooling enrollment at the point of (or soon after) deregulation rather than assuming a gradual linear increase in college enrollment likelihoods since year of deregulation. In particular, if students enroll in college once college loans become available or affordable, one would expect a sudden change to the drop in interest rates that levels off regularly and quickly after deregulation. To show that change, the authors should use a more flexible specification that better maps the timing of interest rate adjustments to support their reduced form specification. A discontinuity following reform would be reassuring: if education jumps in response to an immediate change in interest rates, it is far more convincing that the channel of influence is credit constraints rather than broader changes to the local economic environment.

Aside from various channels through which interest rate changes may influence schooling behavior, a separate, more basic concern is the potential endogeneity of the instrumental variable. That is, is the timing of interest rate deregulation correlated with other trends in local economic or schooling activity? Given that the instrument has been used numerous times in the past, previous papers have presumably dealt extensively with precisely this issue. Nonetheless, it is important to go through the usual exercise of verifying that the timing of banking deregulation *as it varies in this particular analysis* is not correlated with other characteristics of the local economic or schooling environment. In reality, given the age of the sample (respondents were between the ages of 14 and 22 in 1979), this particular analysis makes use of only a small fraction of the total cross-state

variation in deregulation timing, so the variation used to identify schooling responses may be less clean than that used in previous work. Furthermore, only 15 of 48 states plus the District of Columbia had undergone reforms by 1979, and almost all of them did so in 1960, so a continuous variable (years since deregulation took place) is a little misleading. In practice, the variation in timing of interest rate deregulation in 1979 looks much more like a binary variable.

Given those considerations, it would be reassuring to see a direct comparison between "before" and "after" states using variables like local economic conditions and trends in migration and employment around 1979. Although the paper currently relies on previous work to justify the critical assumption that the reforms are unrelated to other economic trends, given that nature of the variation that is specific to this paper, it is important to show that there are no key differences at this particular point in time (1979) between "after" states—those that did not undergo reforms until after 1979—and "before" states—those that underwent reforms in 1960 (nine cases, excluding the District of Columbia) or during the 1970s (six cases, spread over the decade).

With regard to the econometrics, the specification employed by the authors is otherwise very straightforward, but the analysis would benefit from a handful of sensitivity checks. In particular, it would be helpful to show that results are robust to the inclusion of high school dropouts who attain their General Educational Development certificates (GEDs), whose schooling decisions (and possibly the decision to acquire a GED) should also be sensitive to cost considerations. The heterogeneity analysis on which many of the conclusions rest is also subject to the same concern that plagues all subgroup analyses: it may be that differences across income groups proxy for some other characteristic of students that has nothing to do with interest rate sensitivity. To address that concern, household fixed effects would provide a useful check on any potential family background characteristics of this nature. Finding a pattern of results of similar magnitude based on a specification that contains household fixed effects would rule out certain possibilities, such as households with high-ability kids also having more creditworthy parents. It would also be reassuring to verify that the results are robust to the inclusion of any right-hand-side variables that are potentially endogenous to schooling choices, including self-esteem and locus of control scores, which surely vary (within person) with increases in schooling attainment.

REFERENCES

Black, Sandra E., and Philip E. Strahan. 2002. "Entrepreneurship and Bank Credit Availability." *Journal of Finance* 57 (6): 2807–33.

Demyanyk, Yuliya, Charlotte Ostergaard, and Bent E. Sørensen. 2007. "U.S. Banking Deregulation, Small Businesses, and Interstate Insurance of Personal Income." *Journal of Finance* 62 (6): 2763–801.

Huang, Rocco R. 2008. "Evaluating the Real Effect of Bank Branching Deregulation: Comparing Contiguous Counties across U.S. State Borders." *Journal of Financial Economics* 87 (3): 678–705.

Jayaratne, Jith, and Philip E. Strahan. 1996. "The Finance-Growth Nexus: Evidence from Bank Branch Deregulation." *Quarterly Journal of Economics* 111 (3): 639–70.

Kerr, William R., and Ramana Nanda. 2009. "Democratizing Entry: Banking Deregulations, Financing Constraints, and Entrepreneurship." *Journal of Financial Economics* 94 (1): 124–49.

Morgan, Donald P., Bertrand Rime, and Philip E. Strahan. 2004. "Bank Integration and State Business Cycles." *Quarterly Journal of Economics* 119 (4): 1555–85.

Foreign Scientists and Engineers and Economic Growth

Giovanni Peri, Kevin Shih, and Chad Sparber

ABSTRACT

Attracting highly educated immigrants—especially scientists and engineers—is a potentially effective economic growth–promoting strategy. This paper evaluates the contribution of foreign-born scientists and engineers to the wage and employment growth of native-born workers. First, we analyze the effects of an increase in the number of foreign H-1B workers across U.S. cities and Federal Skilled Worker Program immigrants across Canadian cities. Second, we compare the effects of attracting scientists and engineers as a growth strategy against alternatives such as pursuing fast-growing industries and expanding tertiary education institutions. We conclude by arguing that the H-1B program can explain a quarter of the wage growth of U.S. college-educated workers from 1990 to 2010, and that the currently proposed enlargement of the program could generate an additional 2 percentage points of wage growth for highly educated natives over the next 20 years.

Giovanni Peri is professor of economics at the University of California, Davis, and a research associate with the National Bureau of Economic Research. Kevin Shih is a doctoral candidate in economics at the University of California, Davis. Chad Sparber is associate professor of economics at Colgate University.

The authors thank Daniel Shoag and Jeffrey Smith for their comments.

Foreign Scientists and Engineers and Economic Growth

1. INTRODUCTION

Skilled workers are highly mobile across international borders. Their high mobility is in part because many countries have adopted immigration policies that favor skilled immigrants.[1] Ostensibly, such policies are intended to attract investors, entrepreneurs, and workers with scarce talents who can capably generate direct economic benefits for the receiving country and its native populace. An immigrant business owner, for example, might hire local labor and decrease the unemployment rate. Similarly, an increase in the science and engineering work force should lead to increased production of knowledge, creativity, and innovation. When localized in one area (such as a city or region), the agglomeration of skilled individuals and the knowledge creation and diffusion that they promote generate productive spillovers to other industries and sectors, increasing economic activity.[2]

This paper estimates the size of the macroeconomic gains for host countries generated by skilled immigrants by evaluating the role of foreign-born scientists and engineers (S&E) in affecting the wage and employment outcomes of native-born workers. By analyzing together the native employment and wage effect of foreign-born S&E, we are able to characterize whether they enhanced productivity (labor demand) growth or, instead, created competition for native workers. We find strong evidence that such immigrants generated large wage benefits for native-born college-educated workers in the United States and Canada. Wage effects for native workers without a college education, as well as employment effects for all native workers, are mostly statistically insignificant. Those long-run results indicate that foreign-born S&E have contributed to productivity growth, particularly of the "skill-biased" type.

[1] See OECD (2010, 2011, 2012).
[2] See Griliches (1992), Ellison and Glaeser (1999), Jones (1995), Moretti (2004a, 2004b), and Iranzo and Peri (2009) for examples.

Section 2 of this paper provides an overview of international trends in skilled-labor mobility, with a particular emphasis on S&E. It begins with a brief discussion and an overview of international skilled-labor mobility. The subsections describe the skilled-migration policies of two countries that attract a very large share of the world's S&E and will be the focus of our analysis: the U.S. H-1B visa program and the Canadian points-based Federal Skilled Worker Program (FSWP).

Section 3 describes a simple framework to interpret our results and discusses the empirical model. Section 4 presents the data and describes and discusses the identification strategy. We exploit cross-city variation to assess the effect of foreign-born S&E on the wages and employment of native-born workers with high and low levels of education. Importantly, such immigrants do not randomly choose their work locations. Thus, much of Section 4 is devoted to outlining our method of identifying the causal effects. We combine historical cross-city variation of skilled-immigrant populations with national-level skilled-worker (H-1B or FSWP) inflows to create exogenous instruments for a city's foreign-born science and engineering work force. We devote significant time to establishing the validity of our instruments and exploring alternatives to verify the robustness of our methodology.

We perform the primary regression analysis for the United States and Canada in Sections 5 and 6, respectively. An increase in foreign-born S&E by 1 percentage point of employment increased the wage growth of native-born college-educated workers by around 5 percentage points in the United States. The analogue estimate for Canada ranges from 2.8 to 5.5 percentage points. No regression uncovers statistically significant detrimental effects on wages paid to non-college-educated natives or on the employment levels of native workers. The results, interpreted within the framework of our simple labor demand-and-supply model at the local level, imply that foreign S&E have increased the labor demand and productivity of highly educated native-born workers, leaving unchanged the wage and employment of less educated native workers. The skill-biased growth brought by foreign S&E more than offsets the direct competition–complementarity effect from their presence in the labor market.

One interesting difference between the U.S. and Canadian experience does emerge. For the United States, science and engineering immigrants benefit only highly educated natives, suggesting that such

immigrants facilitate skill-biased technological change. For Canada, natives who have dropped out of high school also appear to reap some benefits from skilled immigration (especially in the form of faster employment growth), suggesting that such immigrants might have contributed to the polarization of the labor market (i.e., growth at the high and low ends of the education spectrum at the expenses of intermediate-level jobs), as described by Autor, Katz, and Kearney (2006) and Autor (2010).

Section 7 focuses on the U.S. experience and assesses whether the beneficial effects of immigration remain after controlling for other potential growth-promoting factors, such as efforts to attract high-growth industries. Not only do the immigration results remain, but they also reveal wage-enhancing effects larger in magnitude than those from other skill-promoting policies. Moreover, we find that the strongest positive productivity effects of H-1B-driven scientists and engineers are in metropolitan areas with a large share of fast-growing industries. The section closes by highlighting an important aspect of our analysis: since our identification strategy is driven by changes in the available number of H-1B visas, our regression strategy also serves as an evaluation of the H-1B program. Therefore, we combine our regression results with observed visa data to estimate that H-1B-driven science and engineering labor flows increased the wage growth of native-born college-educated workers by 3.5 percentage points between 1990 and 2010, about a quarter of the observed wage growth over that period. Current proposals to raise annual H-1B issuances by 50,000 per annum would translate to an additional 1.8 percentage points in wage growth rates of native college-educated workers over the next 20 years.

2. INTERNATIONAL COMPETITION FOR S&E

Competition to attract skilled immigrants has intensified in recent years, and it may grow even more in the near future. Australia, Canada, and New Zealand have long used points-based systems (PBSs) where work permits and immigration visas are awarded to applicants who accrue enough points to exceed a predetermined point threshold. Having high educational attainment, being young, speaking the local language, and offering other broadly defined skills confer an advantage to potential immigrants by endowing additional points to applicants with such characteristics. Since 2008,

PBSs have also been used in Austria, Denmark,[3] Japan, the Netherlands, South Korea, and the United Kingdom. Countries without PBS policies but with selective immigration criteria favoring high-skilled immigration include Germany, Lithuania, Malaysia, Norway, Singapore, Switzerland, Taiwan, and the United States (through the H-1B program). Importantly, Asia has traditionally been a major supplier of skilled migrants to countries that are members of the Organization for Economic Cooperation and Development (OECD), and Asia continues to educate large numbers of people. However, Asia's gross domestic product is rising, Asian countries are taking steps to adopt skill-biased immigration policies, and fertility rates in Asia are falling. Those developments may intensify the international competition for skilled workers.

Although the stated goals of skilled-immigration policy consist in attracting particular attributes (e.g., entrepreneurship, investment, scientific capability), countries' immigration policies have used a broad definition of skills that is strongly linked to educational attainment, as opposed to a narrow definition more closely linked with occupations. The U.S. H-1B program requires immigrants to be employed in a job from a list of approved occupations, but that list is wide-ranging. Both Australia and Canada recently dropped occupation-specific point criteria in their PBS programs in favor of points awarded for more general skill sets.

Despite the absence of occupational targets, however, large numbers of migrants work in science and engineering positions because of the inherent nature of such work. Hunt and Gauthier-Loiselle (2010), for example, note that foreign-born workers are likely to be overrepresented in science and engineering occupations because the required knowledge does not rely on institutional or cultural knowledge, nor does it demand the sophisticated language fluency necessary in fields like law. Thus, science and engineering knowledge and skills transfer easily across countries. Similarly, Peri and Sparber (2011) argue that skilled immigrants have a comparative advantage in math and science, whereas natives have a comparative advantage in communication skills.

Reliable cross-country data on S&E and highly educated labor mobility are somewhat limited. Two new data sets provide some

[3] Denmark has since discontinued its PBS.

numerical insight. First, the OECD (2012) and its Database on Immigrants in OECD Countries report that the overall emigration rate for non-OECD countries in Asia is 3.8 percent for the highly educated, but just 0.3 percent for medium- and low-education workers. Second, Franzoni, Scellato, and Stephan (2012) have developed a new "GlobSci" survey data set that records the characteristics of scientists employed in four fields (biology, chemistry, earth and environmental science, and materials) within 16 countries that account for 70 percent of academic articles published in those fields.[4] The source country with the greatest mobility is India; nearly 40 percent of Indian scientists work abroad. Though India is an outlier, emigration rates are high in other countries as well: more than 20 percent of Belgian, British, Canadian, Dutch, German, and Swiss scientists are currently working outside their home countries. For 13 of the countries covered, over half of their native-born scientists have some international work experience, with Italy (40 percent), Japan (39.5 percent), and the United States (19.2 percent) being the three exceptions.

In some ways, highly educated foreign labor is concentrated in just a few countries. In 2005–2006, for example, roughly 60 percent of highly educated Asian migrants in OECD countries lived in the United States, with another 30 percent living in Australia, Canada, and the United Kingdom.[5] The United States accounts for 45 percent of the OECD's highly educated immigrant stock from all regions, followed by Canada (11 percent), the United Kingdom (9 percent), and Germany (6 percent).[6] In proportional terms, other countries are far more reliant on foreign-born labor. In Switzerland, 56.7 percent of scientists are foreign-born, with about a third of those workers coming from Germany. Immigrants make up roughly 45 percent of scientists in Australia and Canada, and about 38 percent in the United States and Sweden.[7]

How much those concentrations of foreign S&E contribute to the productivity of the receiving country is the open question we seek to

[4] China is the most significant destination country omitted from the survey.

[5] Inferred from Table III.1 of OECD (2012).

[6] Data from Database on Immigrants in OECD Countries.

[7] Also see National Science Board (2012); Kerr and Lincoln (2010); Hunter, Oswald, and Charlton (2009); and Auriol (2007, 2010).

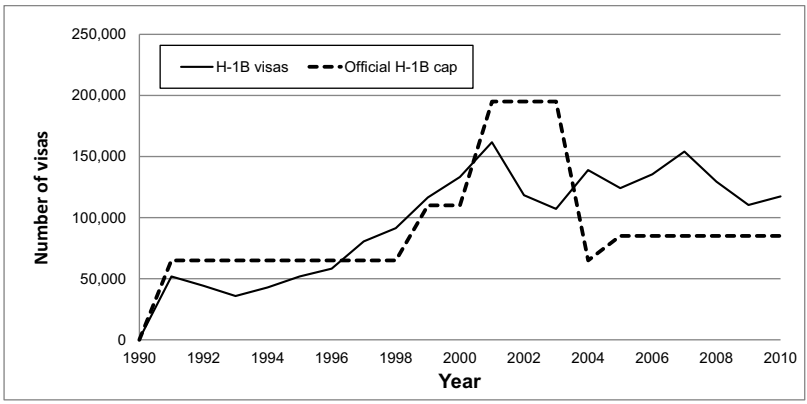
address. Past research on the United States tends to provide evidence of productivity gains from skilled immigration. Kerr and Lincoln (2010) find that H-1B workers increase Indian and Chinese invention in the United States without crowding out natives from the innovation process. Hunt and Gauthier-Loiselle (2010, p. 33) argue that immigrants account for 24 percent of U.S. patents—twice their share of the population—and that the immigrant patenting advantage over natives is accounted for by immigrants' disproportionate tendency to hold degrees in science and engineering fields. Similarly, Hunt (2011, p. 421) finds that immigrants are more entrepreneurial and innovative than natives and that much of that advantage is explained by immigrants' higher education and field of study. Thus, U.S. firms, universities, and teaching hospitals identify innovative immigrants mainly on the basis of their educational qualities rather than on superior innate creative or inventive abilities at a given educational level. Immigrants do, however, exhibit higher publication rates than natives even after controlling for education. Recent work by Scellato, Franzoni, and Stephan (2012, p. 2) complements that result by finding that (a) foreign-born scientists maintain research links with colleagues from their home country and (b) internationally mobile researchers contribute significantly to extending the international scope and quality of the research network in destination countries at no detriment to the quality of the research performed.

2.1 The U.S. H-1B Visa Program

The Immigration Act of 1990 created the H-1B visa for college-educated foreign-born temporary workers in the United States. The H-1B visa is the most common route of entry for temporary skilled workers. U.S. State Department statistics, for example, reveal that 117,409 new H-1B visas were issued in 2010—a figure that is 57 percent larger than the next most important temporary worker program, the L-1 visa for intracompany transferees.

Figure 1 illustrates the evolution of the actual number of H-1B flows and the legal cap on those flows for each year since the program's inception. Congress initially limited the number of new H-1B visas to 65,000 per year. The cap was not reached until 1997 and 1998. In October 1998, Congress raised the cap to 115,000 for 1999 and 2000, but those limits were also reached (or exceeded). Thus, Congress took two actions in October 2000. First, it exempted employees of

Figure 1
Evolution of the H-1B Visa



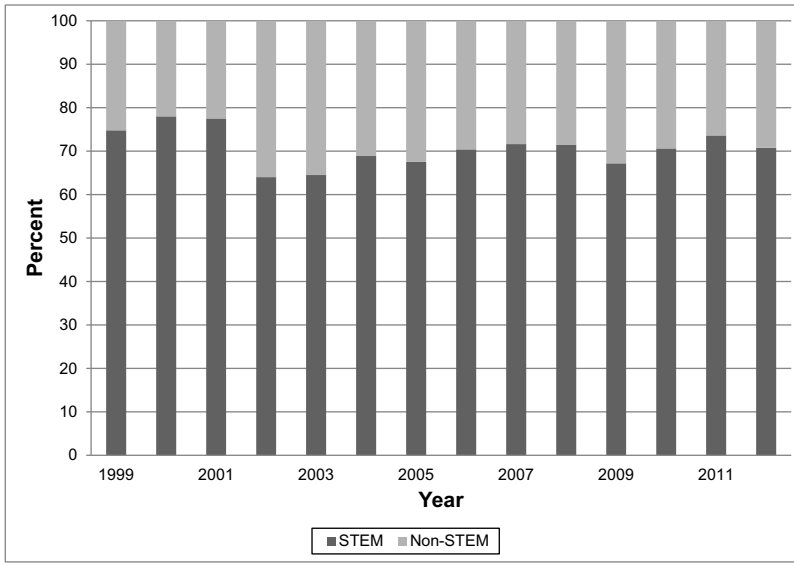
Source: Data on H-1B visas issued available from Department of State, http://www.travel.state.gov/xls/FYs97-12_NIVDetailTable.xls.

universities, nonprofit research organizations, and government research organizations from the H-1B limit. Second, it raised the annual cap to 195,000 for 2001, 2002, and 2003. In 2004, the cap reverted back to the original 65,000, but the exemptions remained. In 2005, Congress began exempting 20,000 visas for workers who have obtained a master's degree or higher in the United States, effectively raising the cap to 85,000. Nonetheless, the H-1B cap has been binding every year since 2004.

The H-1B visa is not limited to S&E, and publicly available data on formal links between H-1B issuances and S&E work are limited at the individual level.[8] However, the U.S. Citizenship and Immigration Services annual reports (1998–2011) on the "Characteristics of H-1B Specialty Occupation Workers" provide aggregate statistics on H-1B petitions granted by occupational category. Figure 2 displays the share of H-1B visas awarded to science, technology, engineering, and mathematics (STEM) workers, who clearly dominate the program. The figure shows that STEM occupations (categorized as

[8] This point is also noted in Kerr and Lincoln (2010).

Figure 2
STEM Workers as a Percentage of H-1B Visa Recipients



Source: Data from U.S. Citizen and Immigration Services reports from 1999–2012 entitled "Characteristics of H-1B Specialty Occupation Workers."
Note: STEM = science, technology, engineering, and mathematics.

computer-related, engineering and architecture, medical and health sciences, life sciences, and math and physical sciences) accounted for about 70 percent of the H-1B visas issued. Computer-related occupations accounted for 47 percent of STEM H-1B visas, while engineers represented almost 12 percent.

Whereas data from the U.S. Department of Homeland Security provide key information on H-1B visa issuances, the U.S. Labor Department's Office of Foreign Labor Certification provides an alternative data set on Labor Condition Applications (LCAs) that provides insights into the nature of temporary H-1B work demanded by U.S. employers. Firms wishing to employ H-1B workers must first submit an LCA describing the available occupation, the prevailing wage, and the wage offer to the foreign worker. The Labor Department can then

approve or deny the request. LCAs, therefore, contain a wealth of information on the types of workers that firms wish to hire.[9]

Data from fiscal year 2011 show that 91 percent of LCAs were filed for science and engineering positions.[10] Among S&E, computer systems analyst (76.5 percent) is the most commonly sought position, followed by computer programmer (10.7 percent). No other occupation exceeds 5 percent of the LCA pool for S&E. Among non-S&E filings, the most common occupations are accountants and auditors (9.7 percent), management analysts (8.2 percent), financial analysts (6.8 percent), and farm labor contractors (6.2 percent).[11] Statistics from the Department of Homeland Security's *Report on Specialty Occupation (H-1B) Workers* (2011) compare with some of those figures. Some 73.6 percent of approved petitions are in STEM occupations, including 50.8 percent for computer-related occupations, 2.4 percent for life sciences, 2.2 percent for math and physical sciences, 7.0 percent for medicine and health, and 11.3 percent for architecture, engineering, and surveying. Overall, those aggregate statistics suggest that it is reasonable to view the H-1B program as being responsible for bringing the bulk of foreign scientists and engineers into the United States between 1990 and 2010.

2.2 The Canadian Points-Based Federal Skilled Worker Program

Canada has been using PBS since 1967, with significant legislative changes occurring with the Immigration Act of 1976 and the Immigration and Refugee Protection Act (IRPA) of 2002. Unlike the U.S. H-1B visa, Canada's FSWP offers foreign-born skilled laborers permanent residency. Like the U.S. H-1B visa, Canada's skilled-immigration

[9] Page 5 of the Department of Labor's 2011 "Annual Report" notes, "Certification of a position, however, is not a guarantee of a foreign worker's admission, since many visa categories . . . have numerical limitations or caps set by legislation, and each individual must meet admission standards and requirements of [the Department of Homeland Security] and [the Department of State]." See Kerr and Lincoln (2010, 485) for further comment.

[10] In the LCA data, the science and engineering class includes science managers and instructors, but not medical doctors or health practitioners.

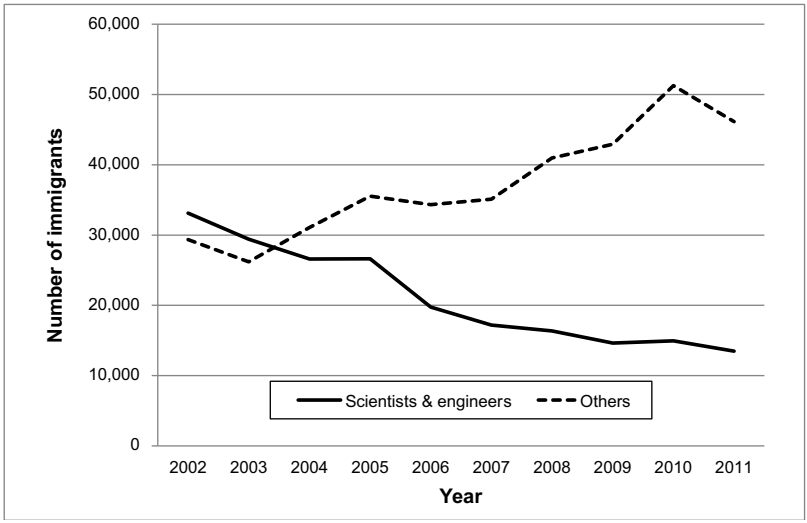
[11] For comparison, Lofstrom and Hayes (2011) used the Freedom of Information Act to obtain data on actual H-1B recipients (in 2009) through I-129 forms. They find a more modest 74 percent of H-1B workers to be employed in science, technology, engineering, and math professions, with only 57 percent of those individuals working in information technology jobs.

policy has become both more and less restrictive at different points over the past decade and a half. For example, the minimum score required for permanent residency was reduced in September 2003, whereas recent years have seen somewhat subtle but perhaps significant restrictions on skilled immigration, ostensibly to attract a foreign-born work force capable of adjusting to Canadian life. Though there has never been a limit on the number of permanent resident visas awarded to qualified skilled immigrants with Canadian job offers, a cap of 20,000 was introduced in 2010 for skilled immigrants without job offers. That cap declined to 10,000 in 2011 and is currently at just 5,000 for the next year. Moreover, Canada suspended applications for immigrants without job offers from July 1, 2012, through May 4, 2013.

Perhaps most important, pre-IRPA criteria favored specific occupations by giving more points to some over others. IRPA, however, abandoned the occupation-specific points scheme and focused on rewarding broader measures of human capital, such as educational attainment, to improve labor-market flexibility among Canada's foreign-born work force. Citizenship and Immigration Canada's (2010) evaluation of those program changes found that IRPA diversified the skilled-immigrant pool. Proportionally fewer immigrants now come from Asia, and the share of skilled immigrants working in the natural and applied sciences fell. Figure 3 illustrates trends since 2002 and demonstrates that it is not just the science and engineering share of foreign-born permanent Canadian residents that has fallen, but the overall number of immigrant S&E as well.

International comparisons of science and engineering reliance are difficult, in part, because different countries use different occupational coding schemes. The OECD, for example, provides information from 2000–2001 and 2005–2006 about Canadian residents working as professionals, technicians, and other assorted professions (excluding managers), whereas it records the number of computer and mathematical scientists; architects and engineers; and life, physical, and social science occupations for the United States. If we define S&E to be PhD recipients in those occupational groups, then the OECD data in 2000–2001 indicate that 52 percent of Canadian S&E were foreign-born, compared with 38 percent for the United States. In 2005–2006, the foreign-born share decreased to 51 percent for Canada and rose to 43 percent for the United States. Those numbers are slightly higher

Figure 3
Permanent Foreign-Born Residents in Canada



Source: Citizenship and Immigration Canada, 2011 facts and figures.

than the foreign-born share of scientists recognized in Franzoni, Scellato, and Stephan's (2012) GlobSci data set (46.9 percent for Canada, 38.4 percent for the United States).

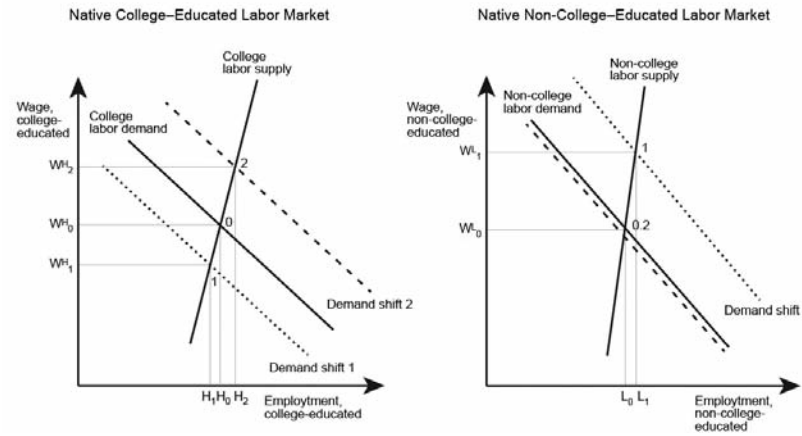
As discussed earlier, many countries have implemented different policies with the goal of attracting highly skilled foreign workers. The United States and Canada have received the largest number of high-skilled immigrant workers by far and thus provide unique settings to study the effect of foreign workers on the receiving economy. In particular, the United States and Canada have targeted high-skilled workers through quite different types of policies. Furthermore, the structure and design of their immigration policies have varied over recent decades. Despite the differences in policies, both countries have seen foreign S&E continually dominate the inflow of foreign high-skilled workers. What has been the effect of inflows of S&E, and does the United States' experience differ from Canada's? We turn to those questions in the next sections.

3. FRAMEWORK FOR THE EMPIRICAL ANALYSIS

The goal of the empirical analysis in this paper is to establish the long-run effects of foreign S&E on the wages and employment of natives. We consider local labor markets for college-educated and non-college-educated workers, and we estimate the effects on the employment and wages of both sets of native workers while using the variation of foreign S&E across metropolitan areas. Using those estimates, we can characterize the effect of foreign S&E on the demand for each native group. Foreign S&E may affect the demand for each group of natives through substitution, complementarity, and productivity effects. We illustrate a simple labor demand-and-supply model for more educated and less educated native workers at the local level to interpret the empirical findings. The model can be fully formalized, and it is similar to the ones presented in Docquier, Odin, and Peri (forthcoming) and Peri (2011).

Consider Figure 4 as representing the demand and supply of native college-educated workers (left panel) and non-college-educated workers (right panel) in one small, open economy, such as a U.S. metropolitan area. An exogenous inflow of foreign S&E into the economy corresponds to a potential shift in the demand for each group (such as those illustrated by the dashed and dotted lines). The demand shift for a native group will be to the left if foreign S&E

Figure 4
Potential Demand Shocks in Labor Market



substitute for native workers of that group. It will be to the right if they complement that type of native workers. Moreover, the productivity of each group can be increased or decreased by the research activity of S&E. They, in fact, contribute to generating innovations and developing technologies that may, in turn, complement or substitute for the abilities of specific workers. Productivity effects, therefore, will also contribute to positive or negative demand shifts. The relative response of native employment and native wages to those shifts will be regulated by the elasticity of the local supply of each type of workers. Lower mobility of workers (or a larger response of local prices) would imply larger wage and smaller employment adjustment (rigid supply). Higher mobility and lack of adjustment of local prices would imply, instead, larger employment and smaller wage adjustment for a given shift in demand (elastic supply). Admittedly, the supply elasticity of a group in our simple model combines moves in and out of employment, the local economy, and education groups (in the long run). Our employment effects do not distinguish across those margins.

To fix ideas, we illustrate two possible effects of a change in foreign S&E on the labor demand for natives, as shown in Figure 4, by the shifts labeled "Demand shift 1" (dotted line) and "Demand shift 2" (dashed line). The first shift produces a negative wage effect for college-educated workers and a positive one for non-college-educated workers. Employment moves in the same direction as wages for each group, but not as far (as the labor supplies are drawn as rather inelastic). Such would be the effect of foreign S&E were they substitutes (competitors) for native college-educated workers, as well as complements for native non-college-educated workers, while having no productivity effect.

A very different shift, however, is shown with the dashed lines marked "Demand shift 2." In that case, the wage of college-educated native workers increases, while the wage and employment of less educated natives remain the same. Such a shift would imply a positive and college-biased productivity effect of S&E that offsets the substitution effect and increases the demand for college-educated natives. It also reveals that the skill bias of productivity is strong enough to offset the positive complementarity with the less educated, and their wages and employment are not affected by the shock.

Estimating the response of employment and wages of natives to an increase in foreign S&E provides an assessment of the demand shock

that each group experienced. It also allows us to evaluate the elasticity of the supply of that group. For instance, a positive wage effect on a group accompanied by a null employment effect implies a positive demand shift for that group and a rather rigid labor supply.[12] With this framework in mind, we describe our empirical specifications, and we later interpret the findings.

3.1 Empirical Specification

Empirically, we analyze the effect of foreign S&E on wages and employment using variation across 219 U.S. metropolitan areas (and later on 17 Canadian areas) over time. For the United States, we use the uneven distribution of H-1B visa holders across nationalities and the pre-policy distribution of foreign S&E by nationality across metro areas to construct a policy-driven increase in S&E. This variable is unaffected by the local productivity shocks and is heterogeneous across metropolitan areas. We then measure the effect of this policy variable on actual foreign S&E, and then on the wage and employment growth of natives across U.S. metropolitan areas.

The basic specifications that we will estimate in Section 5 take the form of the following equations (1) and (2):

$$(1) \quad \frac{\Delta W_{c,t}^{X,Native}}{W_{c,t}^{X}} = \varphi_t + \varphi_c + b_{y,X} \cdot \frac{\Delta(S\&E)_{c,t}^{Foreign}}{Emp_{c,t}} + \gamma \cdot Controls_{c,t}^{X} + \varepsilon_{c,t}$$

$$(2) \quad \frac{\Delta Emp_{c,t}^{X,Native}}{Emp_{c,t}} = \varphi_t + \varphi_c + b_{y,X} \cdot \frac{\Delta(S\&E)_{c,t}^{Foreign}}{Emp_{c,t}} + \gamma \cdot Controls_{c,t}^{X} + \varepsilon_{c,t}$$

The dependent variables are, alternatively, the percentage native wage change (specification [1]) and the change in native workers, over a decade, relative to total initial employment (specification [2]). Each regression is estimated separately for type X, which can be either college-educated (also defined as highly educated) or non-college-educated (also defined as less educated), in city c. By distinguishing workers of different skills (college-educated and non-college-educated), we identify whether the potential productivity, complementarity, and substitution effects of foreign S&E differ across groups of natives.

[12] Considering specifications such as equations (1) and (2), the ratio of the estimated response of employment and of the estimated response of wages to a change in foreign S&E gives the elasticity of the labor supply for that group.

The term $\frac{\Delta(S\&E)_{c,t}^{Foreign}}{Emp_{c,t}}$ is the main explanatory variable, and it measures the change of foreign S&E, $\Delta(S\&E)_{c,t}^{Foreign}$, expressed as percentage points of $Emp_{c,t}$, the total employment in the metro area at the beginning of the period. The coefficient of interest is $b_{y,X}$. In equation (1), it captures the response of native wages, in percentage points, to an increase of foreign S&E by 1 percentage point of total employment. In equation (2), it captures the response of the group employment, as percentage points of total initial employment.

In estimating equations (1) and (2), we are concerned about the potential for omitted variables bias in generating spurious results. Our decision to first-difference the data helps account for time-invariant features correlated with the level of foreign S&E wages and employment across cities. We then include fixed effects (φ_c, for either states or metropolitan areas) and time period effects (φ_t,) to capture features correlated with growth rates across space and time. Further, we add a set of $Controls_{c,t}^{X}$, representing other metro-area factors potentially affecting local labor demand, such as the 1980 share of native S&E, or the 1980 share of college-educated, or the effect of the industrial composition on wages and employment growth at the metro-area level. $\varepsilon_{c,t}$ is a zero-mean random error.

If foreign S&E were assigned exogenously across U.S. metropolitan areas, after accounting for fixed effects and other controls, the ordinary least squares (OLS) estimates of equations (1) and (2) would capture the causal effect of that group on native labor-market outcomes. However, unobserved changes in productivity realized over time in specific metropolitan areas might still be present and might generate a spurious correlation that would bias the estimates. To address those issues, we construct a policy-driven change in foreign S&E. In particular, we use the changes in H-1B numbers by nationality across decades at the U.S. level (which are hardly affected by individual city demand conditions as, except for the top 5 percent, each metro areas accounts for less than 1 percent of national science and engineering employment)[13] and the pre-1980 distribution of foreign S&E across U.S. metropolitan areas (which are unaffected

[13] The top four cities—New York, Los Angeles, San Francisco, and Washington, D.C.—attracted between 4 and 10 percent of the total national foreign S&E in the 1990s and 2000s. We run a specification in Table 6 omitting them, and the results are unchanged.

by demand shocks in the 1990s and beyond) to construct the policy-driven variable $\frac{\Delta(S\&E)_{c,t}^{H\text{-}1B}}{Emp_{c,t}}$. This variable captures changes in foreign S&E driven by national-level fluctuations in H-1B flows and the existence of stronger or weaker networks of immigrants in metro areas as of 1980. We discuss in detail the identification assumptions implied by this variable and some tests of their plausibility in Section 4.3.

4. DATA ON U.S. METROPOLITAN AREAS AND THE POLICY VARIABLE

Our data on occupations, employment, wages, age, and education of individuals come from the Integrated Public Use Microdata Series (IPUMS) 5 percent census files for 1980, 1990, and 2000. We also use the 1 percent sample for 2005 and the 2008–2010 three-year samples of the American Community Survey to obtain a 3 percent sample that we call "2010." We only use data on 219 metropolitan areas that can be consistently identified over the full 1980–2010 period.[14] Those metropolitan areas include all the largest metropolises in the United States (Los Angeles, New York, Chicago, Dallas–Fort Worth, Philadelphia, and Houston are the six largest) down to metropolitan areas with close to 200,000 people (Danville, Va.; Decatur, Ill.; Sharon, Pa.; Waterbury, Conn.; Muncie, Ind.; and Alexandria, Pa. are the six smallest). Data on aggregate H-1B flows by nationality and year, which we use to construct our policy variable, are publicly available from the U.S. Department of State (2011).

Although the U.S. government recognizes a list of official STEM college degrees for the purpose of permitting foreign students to work under the Optional Practical Training program, there is no official definition of science and engineering occupations. We consider two alternative definition criteria. The first is based on the skills used within an occupation. We use the Occupational Information Network (O*NET) database provided by the U.S. Bureau of Labor Statistics, which provides Standard Occupational Classification measures of the importance of several dozen skills and abilities required to perform the job. We select four skills: (a) mathematics in problem solving, (b) science in problem solving, (c) technology design, and

[14] In a robustness check, we will limit the analysis to using data since 1970 for 116 identifiable metropolitan areas.

(d) programming. We consider the average score for each occupation across the four skills and rank the 333 occupations that are identified consistently in the 1980–2010 censuses according to those average scores. We categorize science and engineering occupations as those in the highest decile of STEM skills used by employees in 2000, and we call individuals in those occupations "O*NET S&E." The list of occupations included in this definition is reported in Table A1, Part A, of the Appendix.[15]

The second definition of S&E that we use is also based on occupations, and it classifies them according to the percentage of workers who have obtained college degrees in science and engineering majors as identified by the American Community Survey of 2010.[16] This second definition identifies science and engineering jobs—listed in Table A1, Part B, of the Appendix—as those occupations in which at least 25 percent of workers have graduated with a science or engineering major. This definition is more stringent than the first and comprises about 5 percent of the labor force.[17]

4.1 Aggregate Statistics on S&E

Even a cursory look at the data shows that foreign-born individuals are overrepresented among S&E.[18] Moreover, foreigners have substantially contributed to the aggregate growth of science and engineering jobs in the United States. Table 1 shows the foreign-born share of five different employment groups for each census year from 1980 to 2010. From left to right, we show the percentage of foreign-born among all workers, among college-educated workers, among college-educated workers in metropolitan areas, among science and engineering (O*NET-based) occupations in metropolitan areas, and among college-educated S&E in metropolitan areas. Although foreign-born individuals represented 16.3 percent of total employment in 2010, they counted for a quarter of college-educated S&E in the metropolitan sample that we analyze. Remarkably, that figure has more than doubled since 1980.

[15] The aggregate summary statistics for this definition are shown in Table 1. They are sometimes restricted to those workers with college degrees only.

[16] The science and engineering majors are listed in Table A1, Part C, of the Appendix.

[17] The correlation between the S&E dummies defined for each occupation across the two definitions is 0.45.

[18] In the summary statistics and in the empirical analysis, we use the O*NET S&E definition, unless we note otherwise.

Table 1
Summary Statistics: Percentage of Foreign-Born Groups, United States

	Foreign-Born % of Employment	Foreign-Born % of College-Educated	Foreign-Born % of College-Educated in 219 Metro Areas	Foreign-Born % of S&E in Metro Areas	Foreign-Born % of College-Educated S&E in Metro Areas
1980	6.1	6.7	8.9	9.8	10.9
1990	8.7	8.9	11.9	13.8	14.8
2000	13.2	12.7	16.3	19.6	21.0
2005	15.3	14.7	18.8	23.0	24.4
2010	16.3	15.4	19.4	24.0	25.5

Note: The figures were obtained by the authors from IPUMS census data. The relevant population includes only noninstitutionalized individuals between ages 18 and 65 who have worked at least one week in the previous year.

Table 2 shows that college-educated S&E increased from 2.8 percent of total employment in 1980 to 4.6 percent in 2010. Even more remarkable, the share of college-educated foreign-born S&E in employment grew from 0.3 percent to 1.2 percent. The 1990s were a period of particularly fast growth in S&E jobs relative to other

Table 2
College-Educated O*NET S&E as a Percentage of Total Employment (219 U.S. metropolitan areas)

	Foreign S&E	Total S&E
1980	0.3	2.8
1990	0.5	3.3
2000	0.9	4.5
2005	1.1	4.4
2010	1.2	4.6

Note: The figures were obtained by the authors from IPUMS census data. The relevant population includes only noninstitutionalized individuals between ages 18 and 65 who have worked at least one week in the previous year.

decades in the analysis, as the science and engineering percentage of employment grew by 1.2 points. Of that increase, 0.4 points were due to foreign S&E. Also remarkable, the first decade of the 2000s saw an increase in total S&E by only 0.1 percent of employment, whereas foreign-born S&E increased by 0.2 percent of employment.

Table 3 shows absolute numbers (in thousands), suggesting that the H-1B program was large enough to drive all or most of the increase in foreign college-educated S&E. Column (1) reports the net total increase in college-educated S&E in the United States, and column (2) displays the increase in foreign college-educated S&E. Column (3) shows the cumulative number of H-1B visas issued during the corresponding decade. It is clear that in the 1990s, the H-1B visas were enough to cover the whole growth in college-educated foreign S&E in the United States, even accounting for some returnees. Even more remarkable, H-1B issuances were three times as large as the net increase in college-educated S&E between 2000 and 2010. That increase implies that many foreign S&E, including H-1B recipients, must have left the United States, while many native S&E must have retired, lost their jobs, or changed occupations. Overall, the figures presented emphasize the importance of foreigners for science and engineering jobs in the United States. Foreign-born labor is over-represented among S&E, and the overall size of the H-1B program was large enough to contribute substantially to the foreign science and engineering job growth between 1990 and 2010.

Table 3
Net Increase in College-Educated S&E and Cumulative H-1B Visas (Thousands)

	(1) Net Change in Total College-Educated S&E	(2) Net Change in Foreign College-Educated S&E	(3) Cumulative H-1B Visas
1980–1990	793	196	0
1990–2000	1779	547	707
2000–2005	281	221	650
2005–2010	228	110	647

Note: Data on the change in total S&E occupations are from the IPUMS Census. Data on the total number of H-1B visas are from the U.S. Department of State (2011).

4.2 The Constructed H-1B Policy Variable

We begin by defining for each metro area (city), c, its employment share of foreign S&E from each of 14 specific foreign nationalities,[19] n, in 1980 as $\frac{\Delta(S\&E)^{FOR_n}_{c,1980}}{Emp_{c,1980}}$. The overall foreign S&E employment share in a metro area is the sum of the shares from each specific nationality, $\frac{\Delta(S\&E)^{FOR}_{c,1980}}{Emp_{c,1980}} = \sum_{n=1,14}\left(\frac{\Delta(S\&E)^{FOR_n}_{c,1980}}{Emp_{c,1980}}\right)$. We choose 1980 as a base year for three reasons. First, it is the earliest census that allows the identification of the 219 metropolitan areas that we can consistently follow to year 2010. Second, it is well before the creation of the H-1B visa program, and hence it does not reflect the distribution of foreign S&E produced by the policy. Third, it is early in the information technology (IT) revolution so that the distribution of S&E was hardly affected by the geographic location of the computer and software industries that could be correlated with subsequent positive changes in productivity.[20] Instead, the nuclear, military, chemical, and traditional manufacturing sectors were demanding a large number of science and technology workers in 1980. We also use 1970 as the initial year to determine the foreign S&E distribution for a subset of cities as a robustness check. The instrument based on the 1970 distribution, however, is significantly weaker than the one based on the 1980 distribution.

After defining the 1980 foreign science and engineering employment share, we calculate the growth factor of foreign S&E for each nationality n in the United States between 1980 and year t. We do so by adding the total countrywide inflow of S&E from each nationality during the period between 1980 and t to its initial 1980 countrywide level $(S\&E)^{FOR_n}_{c,1980}$. For the decades 1990–2000 and 2000–2010, we use the cumulative H-1B visas allocated to nationality n $\left(\#ofH\text{-}1B^{FORn}_{1980-1990}\right)$ to proxy for the net aggregate increase

[19] The national groups are from Canada, Mexico, the rest of the Americas (excluding the United States), western Europe, eastern Europe, China, Japan, Korea, the Philippines, India, the rest of Asia, Africa, Oceania, and other.

[20] Although early video games and computers were introduced in the late 1970s, the IBM personal computer was not introduced until 1981.

in $(S\&E)^{FOR_n}$.[21] For the decade 1980–1990, we simply add the net increase in S&E from nationality n as recorded in the U.S. census, $\Delta(S\&E)^{FOR_n}_{1980–1990}$. The imputed growth factor for S&E of each foreign nationality in year t = 1990, 2000, 2005, 2010 is therefore

$$(3) \qquad \frac{\widehat{(S\&E)}^{FOR_n}_{t}}{(S\&E)^{FOR_n}_{1980}} = \frac{(S\&E)^{FOR_n}_{1980} + \Delta(S\&E)^{FOR_n}_{1980–1990} + \#ofH\text{-}1B^{FOR_n}_{1990–t}}{(S\&E)^{FOR_n}_{1980}}$$

To impute the number of foreign S&E in city c in year t, we then multiply the growth factor calculated above for each nationality by the number of foreign S&E of that nationality in 1980 in the city, and then we sum those figures across all nationalities within each city:

$$(4) \qquad \widehat{(S\&E)}^{FOR}_{c,t} = \sum_{n=1,14} (S\&E)^{FOR_n}_{c,1980} \left(\frac{\widehat{(S\&E)}^{FOR_n}_{t}}{(S\&E)^{FOR_n}_{1980}} \right)$$

The H-1B-driven change in foreign S&E that we use as our explanatory variable in the main empirical specifications is the change in $\widehat{(S\&E)}^{FOR}_{c,t}$ over a decade, as percentage points of the initial employment in the city, $Emp_{c,t}$:[22]

$$(5) \qquad \frac{\Delta(S\&E)^{H\text{-}1B}_{c,t}}{Emp_{c,t}} = \frac{\widehat{(S\&E)}^{FOR}_{c,t+10} - \widehat{(S\&E)}^{FOR}_{c,t}}{Emp_{c,t}}$$

This identification strategy is closely related to the one used by Altonji and Card (1991) and Card (2001) and is based on the initial 1980 distribution of foreign workers across U.S. cities. It is also similar to the

[21] Since the data on visas issued by nationality begin in 1997, while we know the total number of visas in each year, we must estimate $\#\text{ofH-1B}_{n,1990–t}$, the total number of visas issued by nationality between 1990 and 1997, as

$$\#ofH\text{-}1B_{n,1990–t} = \#ofH\text{-}1B_{1990–t} \left(\frac{\#ofH\text{-}1B_{n,1990–2010}}{\#ofH\text{-}1B_{n,1990–2010}} \right)$$

where $\frac{\#ofH\text{-}1B_{n,1990–2010}}{\#ofH\text{-}1B_{n,1990–2010}}$ is the share of visas issued to nationality group n among the total visas issued from 1997 to 2010. For t larger than 1997, we have the actual number of yearly visas by nationality $\#ofH1B_{n,t}$.

[22] To avoid the possibility that endogenous changes in total employment at the city level affect the standardization, we also use the imputed city employment (obtained using employment in 1980) augmented by the growth factor of national total employment. Hence, $Emp_{c,t} = Emp_{c,1980} \left(Emp^{US}_{t} / Emp^{US}_{1980} \right)$.

one used by Kerr and Lincoln (2010), who consider the employment share of foreign S&E in an initial year and the effect of H-1B on subsequent innovation. Our variable, however, is based on foreign S&E in a city in 1980 or 1970 (rather than in 1990 as done by Kerr and Lincoln [2010]) and uses the distribution of foreign S&E across 14 nationalities rather than only an aggregate measure. Hence, it should be less subject to correlation with recent economic conditions and should have stronger power if immigrant networks are country specific.[23]

As is clear from Section 2.2, not all foreign S&E enter on H-1B visas, as there are workers entering with other visas or with permanent residency. Moreover, some of the H-1B workers return to their home countries after six years. Hence, as a first step, we need to establish whether our policy-driven variable has predictive power on $\frac{\Delta S\&E_{c,t}^{FOR}}{Emp_{c,t}}$, the observed change in foreign S&E workers in a metropolitan area.

4.3 Identification and Power of the Policy Instrument

Our identification strategy is based on the assumption that a city's employment share of foreign S&E in 1980 was due to the differential presence of immigrants caused by persistent agglomeration of foreign communities. Those differences subsequently affected the supply of foreign-born S&E, but they were not otherwise correlated with future technological and demand shocks that affected wages and employment. A challenge to this assumption is that a city's employment share of foreign S&E in 1980 may be correlated with its productive and industrial structure, specifically with regard to its sector composition and its scientific and technological base in 1980, which, in turn, may subsequently affect future productivity growth.

[23] Note that aggregate flows of H-1B workers at the national level are often constrained (and determined) by changes in federal policy, but they sometimes reflect other conditions occurring in the United States and in origin countries. The fluctuations do not violate exogeneity conditions required for the validity of our instrument so long as they are uncorrelated with individual city-level shocks. This is a reasonable and commonly held assumption used in a number of papers since Altonji and Card (1991). In a robustness check in Table 6, we also check that the few individual metro areas that contribute more than 5 percent to the national number of foreign S&E are not driving the results.

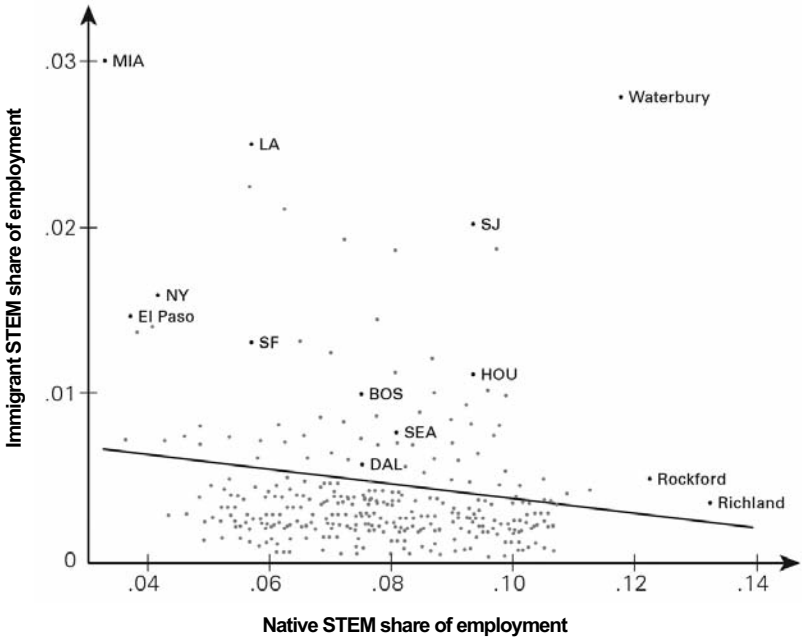
We take several steps to partially address these concerns. First, in this section, we show that the employment share of foreign S&E in 1980 across metropolitan areas has little correlation with the employment share of native S&E, at least when we use the O*NET-based definition. Foreign S&E as a share of employment were much more strongly affected by a city's foreign-born residents. We also show that the share of foreign S&E in 1980 is correlated with the H-1B-driven growth in S&E between 1990 and 2010, independent of the presence of native S&E.

Second, we estimate very demanding empirical specifications. Our data set is composed of a panel of 219 metropolitan areas from 1990–2000, 2000–2005, and 2005–2010. Our models are estimated in first differences so that effects are identified by changes in H-1B-driven foreign S&E supply across cities. That strategy should eliminate bias arising from unobserved time-invariant determinants of the level of our outcome variables and the level of foreign S&E. We add rigor to the models by including controls for 50 state-specific effects and period effects. Thus, identification relies on variation of growth rates across cities in the same state. In robustness checks, we also estimate the more demanding specifications that include 219 city-specific effects. In all specifications, we include controls for the 1980 employment share of native S&E, which proxies for the research intensity of the metro area as of 1980.

And third, we perform robustness checks using the foreign S&E share of employment in 1970 to construct the instrument, and we include other time-varying controls for metro area–specific variables in the regression. In particular, Section 7.1 explicitly controls for the effect of sector composition in 1980 on subsequent growth and analyzes how it affected wage and employment growth in metropolitan areas in comparison with foreign S&E.

The employment share of native and foreign S&E varied dramatically across the 219 U.S. metropolitan areas in 1980, and those two variables had little correlation with each other. Several of the top native science and engineering cities were in the Midwest and the East. Most were associated with traditional sectors—such as nuclear energy, oil products, and industrial machinery—that attracted many S&E in the 1970s. In contrast, the metropolitan areas with large foreign science and engineering employment shares included a more diverse group of cities. Some were larger and had more diversified economies; others were smaller centers with large immigrant communities.

Figure 5
Foreign and Native S&E Employment Shares
(U.S. metro areas, 1980)



Note: The foreign and native S&E employment shares are calculated using 1980 census data for 219 metropolitan areas. The coefficient of the best-fit line = −0.045 (standard error = 0.029). Points are labeled for Boston (BOS), Dallas (DAL), Miami (MIA), New York (NY), Los Angeles (LA), San Francisco (SF), San Jose (SJ), Seattle (SEA), Houston (HOU), El Paso (Texas), Rockford (Ill.), Richland (Wash.), and Waterbury (Conn.).

Figure 5 and column (1) of Table 4 show no correlation between foreign and native science and engineering employment share across cities, when we choose the O*NET-based definition. The OLS correlation obtained after controlling for state effects (column [1]) is negative and insignificant at standard confidence levels (*t*-statistics are smaller than 1.6). The visual impression of Figure 5 is also clear: there was little correlation between foreign and native S&E employment shares in 1980. That fact is a hint that foreign science and engineering

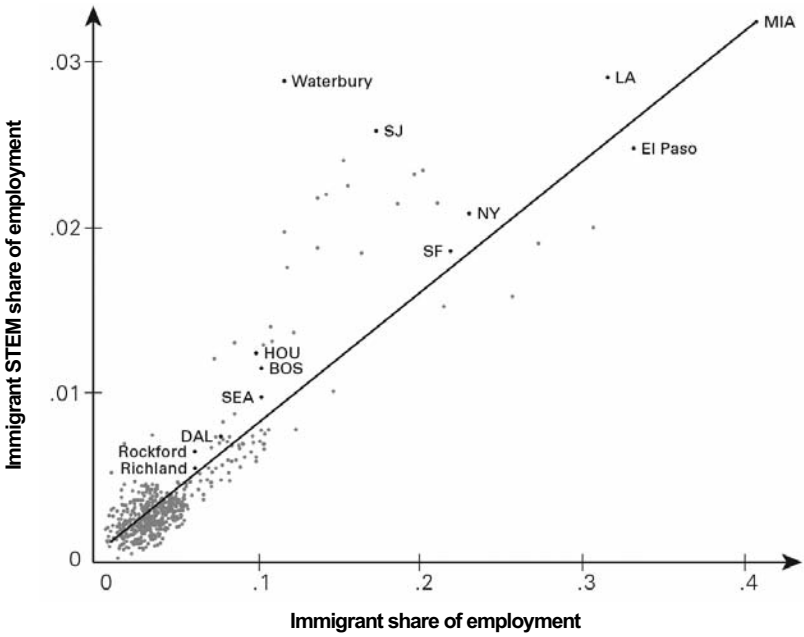
Table 4
Native and Foreign S&E as Shares of Employment in 1980 and the H-1B Predicted S&E Change
(219 U.S. metropolitan areas; 1990–2000, 2000–2005, and 2005–2010)

	(1)	(2)	(3)	(4)
	Foreign S&E as a Share of Employment, 1980	**Foreign S&E as a Share of Employment, 1980**	**H-1B-Predicted S&E Growth**	**H-1B-Predicted S&E Growth**
Foreign-S&E as a share of employment, 1980			0.54*** (0.12)	
Native S&E as a share of employment, 1980	−0.031 (0.032)			0.040** (0.019)
Foreign-born share of population, 1980		0.067*** (0.0065)		
Observations	219	219	657	657
F-statistic	0.94	106.64	20.3	4.38
Period effects	No	No	Yes	Yes
State effects	Yes	Yes	Yes	Yes
Partial *R*-square "Partialing out" state and period effects	N/A	N/A	0.39	0.03

Note: Each column represents a separate regression. The dependent variable is written at the top of the corresponding column. Specifications (1) and (2) include 219 metropolitan areas in 1980. Regressions (3) and (4) include the H-1B-predicted change in S&E in 1990–2000, 2000–2005, and 2005–2010 regressed on the 1980 S&E dependence (foreign or native). The standard errors are heteroskedasticity-robust, and when there is more than one observation per metro area, they are clustered at the metro-area level. The S&E definition is based on O*NET skills. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively. N/A = not applicable.

had little to do with the S&E employment share of a city in 1980. Column (2) of Table 4 and Figure 6 show that a city's employment share of foreign S&E in 1980 had more to do with the presence of other foreign-born residents as a share of the population. When including state fixed effects, the foreign-born population share has an extremely significant association with its foreign S&E employment share (*t*-statistic of 10.3). Anecdotally, cities such as Miami, Los Angeles, and El Paso, Texas, had a large presence of both foreign

Figure 6
Foreign S&E Employment Share and Foreign-Born Employment Share (U.S. metro areas, 1980)



Note: The figures are calculated using 1980 census data. The population of reference used to calculate the foreign-born share of S&E in a city is the total adult (ages 18–65) noninstitutionalized population. The coefficient of the best-fit line = 0.071 (standard error = 0.006). Points are labeled for Boston (BOS), Dallas (DAL), Miami (MIA), New York (NY), Los Angeles (LA), San Francisco (SF), San Jose (SJ), Seattle (SEA), Houston (HOU), El Paso (Texas), Rockford (Ill.), Richland (Wash.), and Waterbury (Conn.).

S&E and foreign workers overall in 1980. Their native-born work force, in contrast, had few S&E.[24] San Francisco, New York, and Los Angeles had a relatively high share of foreign S&E in 1980 but a very unremarkable share of native S&E. Those cities had a large proportion of immigrants, however, and their subsequent growth of S&E has been quite large.

Column (3) of Table 4 goes on to show that the 1980 foreign S&E employment share has extremely significant power to predict the H-1B-driven increase in S&E across cities. (It has an *F*-statistic of 20.30, and the partial *R*-squared explained by that variable is 0.39.) Column (4) of Table 4 shows instead that the 1980 native S&E employment share has very limited power to predict the H-1B-driven increase in S&E (*F*-statistic of 4.38 and partial *R*-squared of 0.03). Cities with larger foreign S&E shares in 1980 were not necessarily associated with a high share of S&E overall in 1980. Even controlling for their native S&E, the fact that the H-1B program allowed a significant increase in the highly educated foreign science and engineering work force during the 1990s and 2000s enabled those cities to increase the size of their science and engineering employment. The initial advantage in employment share of foreign S&E made those cities a more likely destination for foreign-born S&E entering with an H-1B visa. The presence of a network, the easier diffusion of information across foreign groups, and the familiarity of firms with foreign S&E likely reduced the costs for U.S. companies to connect with foreign S&E and for H-1B visa recipients to locate in those cities.

Finally, let us emphasize that our identification strategy relies on more than just the overall foreign S&E share. Since we consider H-1B visas by nationality, we exploit variation in the immigrant presence across U.S. cities by nationality. A very large share of H-1B visas were awarded to Indian, Chinese, and other Asian workers (see Table A2 in the Appendix). Hence, aggregate flows of S&E from those countries more significantly affected cities with large employment shares of S&E from those nationality groups.

[24] Table 4 and Figures 5 and 6 use our first (O*NET-based) definition of science and engineering occupations. Our second definition of those occupations (based on the majors of college graduates) does exhibit correlation between foreign and native S&E shares across cities in 1980. That is why in all regressions we control for the 1980 native S&E share of employment in order to control for the initial science and technology intensity of the metro area.

The H-1B-driven increase in S&E, defined in expression (5), is the policy instrument. It captures the potential effect of the aggregate H-1B visas on S&E in U.S. cities. First, we want to establish that it significantly affected the actual increase in foreign S&E across cities. Then, we would like to determine the effect of S&E on native employment and wages, and hence we use the policy variable as an instrument for the actual increase.

We estimate the following specification:

$$\frac{\Delta(S\&E)_{c,t}^{FOR}}{Emp_{c,t}} = \varphi_t + \varphi_s + b_1 \frac{\Delta(S\&E)_{c,t}^{H\text{-}1B}}{Emp_{c,t}} + \varepsilon_{c,t} \tag{6}$$

The coefficient of interest is b_1, which measures the effect of H-1B-driven S&E inflows on the actual increase in foreign S&E (as measured from the U.S. census). The term φ_t captures period fixed effects, and φ_s represents local area effects (we will include different effects, state or metro-area level, in alternative specifications). We include t = 1990, 2000, 2005 so that the changes Δ refer to the periods 1990–2000, 2000–2005, and 2000–2010.[25] $\varepsilon_{c,t}$ is a zero-mean random error uncorrelated with the explanatory variable.

In Table 5, we show estimates of the coefficient b_1 from different specifications and samples. They illustrate the robustness of the H-1B policy instrument in predicting actual changes in foreign S&E across U.S. metropolitan areas. Columns (1) and (2) show the estimates of the coefficient b_1 in equation (6). In column (1), we include the time dummies and state fixed effects. In column (2), we include the very demanding metro-area fixed effects. The effect of H-1B-driven S&E is always significant at the 5 percent level, and in the basic specification it is close to 0.7, implying that a 1 percentage point increase in the growth of H-1B-driven foreign S&E standardized by initial employment (i.e., the growth rate of foreign S&E) produces an actual increase in the growth rate of foreign S&E in a city by 0.7 percentage points. We can interpret this regression as the first stage in a two-stage least squares (2SLS) estimate of the effect of science and engineering workers. Note that the F-statistic of 17.86 in the basic specification is well above the critical value for the weak instruments test. Only when we include city effects, in column (2), does the policy-driven

[25] We also estimate a specification that uses only decade changes 1990–2000 and 2000–2010, reported in the fifth row of Table 6. The results are very similar to the basic specification.

Table 5

Power of the H-1B Policy Variable as a Predictor of Foreign-S&E Change
(219 U.S. metropolitan areas; 1990–2000, 2000–2005, and 2005–2010)

Dependent Variable	(1) Change in Foreign S&E as a % of Initial Employment (O*NET def.)	(2) As (1) with MSA Effects (O*NET def.)	(3) Change in Total S&E as a % of Initial Employment (O*NET def.)	(4) As (1), including 1980 Share of College-Educated as a Control	(5) As (2) Using Major-Based S&E Definition and MSA Effects	(6) As (5) omitting the Outliers (San Jose, Calif., and Stamford, Conn.)	(7) IV Constructed Using 1970 Foreign Major-Based S&E (116 Metro Areas)
H-1B Driven Growth in Foreign-S&E	0.70*** (0.17)	2.62** (1.19)	0.87** (0.42)	0.63** (0.21)	3.56*** (0.80)	2.68*** (0.49)	0.29* (0.17)
1980 Employment Share of Native S&E	−0.03 (0.02)	N/A	−0.09 (0.08)	−0.02 (0.02)	N/A	N/A	N/A
Observations	657	657	657	657	657	651	348
Period Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Effects	Yes	No	Yes	Yes	No	No	Yes
Metro-Area Effects	No	Yes	No	No	Yes	Yes	No
Other Controls	No	No	No	1980 college-educated share of population	No	No	No
F-test of the Coefficient	17.86	4.85	4.34	9.25	20.02	29.85	2.96

Note: IV = instrumental variable; MSA = metropolitan statistical area; N/A = not applicable. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

variable (though still significant) become less powerful in predicting actual foreign S&E growth (*F*-statistic equal to 4.85).

In column (3) of Table 5, we explore whether the H-1B policy-driven variable had a significant effect on the total increase in S&E. Although less powerful than in predicting foreign S&E growth, the H-1B policy variable has a significant effect (at the 5 percent level) on the growth of total S&E (always standardized by initial employment). That effect implies that more H-1B workers in a city produced more S&E overall.

In columns (4)–(7) of Table 5, we perform several robustness checks of the basic specification (6). Column (4) shows the power of the H-1B-driven growth on foreign S&E when we control for the initial share of college-educated workers in the city. Although the power is somewhat reduced, it is clear that our instrument captures much more than a simple proxy for initial human capital intensity of the metro area. In specification (5), we use the alternative definitions of science and engineering occupations according to college major, and we include occupations whose share of total workers with a science or engineering degree is at least 25 percent or more. We also include the very demanding metropolitan area effects. We find that the policy variable is a very strong predictor of foreign S&E growth. The college major–based definition shows a very strong predictive power: the *F*-statistic remains high (20.02), even with the demanding inclusion of metro-area fixed effects in column (5). That result is reasonable since H-1B visa policy generally restricts admission to highly educated foreign workers, who mainly work in highly science- and technology-intensive occupations, as described in Section 2.2.

In column (6), we omit from the analysis two clear outlier cities whose foreign S&E growth as share of employment has been much larger than in any other city. Those cities are San Jose, Calif., the core of Silicon Valley and its computer industry; and Stamford, Conn., a service center for both technology and financial services. The predictive power of the instrument becomes even stronger, implying that the extraordinary success of Stamford and San Jose in attracting foreign S&E could not be fully predicted by their share of foreign S&E in 1980.

Finally column (7) addresses the possibility that the 1980 foreign S&E distribution might have been influenced by the presence of early IT industries, which, in turn, might have driven productivity in the 1990–2010 period. We construct the H-1B-driven variable using the foreign S&E employment share of cities as revealed by the 1970 census, a year in which computer and information technologies were in their very

early infancy. Unfortunately, this approach restricts our analysis to just the 116 metropolitan areas that can be consistently identified for the whole 1970–2010 period. The power is drastically reduced (F-statistic of 2.96); however, we still find that the H-1B-driven variable significantly predicts the foreign S&E growth in the 1990–2010 period.

5. THE EFFECT OF FOREIGN S&E ON NATIVE WAGES AND EMPLOYMENT

Our H-1B-driven variable has significant predictive power, even after controlling for the technological intensity of metropolitan areas in 1980 and several other observed and unobserved features of those metro areas. Using this source of exogenous variation of foreign S&E, we now analyze their effects.

5.1 Basic Specification

In Table 6 we report the estimated values of the coefficient $b_{y,X}$ from regressions specified in equations (1) and (2). Each of the four columns represents a different outcome. In column (1), the dependent variable is the percentage change of the weekly wage paid to native college-educated workers, $\frac{\Delta w_{c,t}^{College,native}}{w_{c,t}^{College,native}}$, in each of 219 metropolitan areas over the 1990–2000, 2000–2005, and 2005–2010 periods. In column (2), the dependent variable is the percentage change of the weekly wage of native non-college-educated workers, $\frac{\Delta w_{c,t}^{NoCollege,native}}{w_{c,t}^{NoCollege,native}}$.[26] Columns (3) and (4) show the effect of foreign S&E on the change in employment of native college-educated workers and native non-college-educated workers as a percentage of initial total employment (respectively, $\frac{\Delta H_{c,t}^{nat}}{Emp_{c,t}}$ and $\frac{\Delta L_{c,t}^{nat}}{Emp_{c,t}}$).[27]

[26] Weekly wages are defined as yearly wage income divided by the number of weeks worked. Employment includes all individuals ages 18–65 who have worked at least one week during the previous year and do not live in group quarters. Individual weekly wages are weighted by the personal weight in the census. We convert all wages to current 2010 prices using the consumer price index deflator provided by IPUMS.

[27] Column (5) of Table 6 shows the Kleinberger-Paap Wald F-statistic for the first-stage regression to give a sense of the strength of the instruments (essentially identical for all the regressions in the row, as the first stage is the same). A value of 10 is considered a threshold above which issues of weak instruments should not arise.

Table 6

Effects of Foreign S&E on Wage and Employment Growth of Native Workers (219 U.S. metropolitan areas; 1990–2000, 2000–2005, and 2005–2010)

Explanatory Variable: Growth Rate of Foreign S&E **Instrument:** H-1B Imputed Growth of Foreign-S&E	**(1)** **Weekly Wage, Native College-Educated**	**(2)** **Weekly Wage, Native Non-College-Educated**	**(3)** **Employment, Native College-Educated**	**(4)** **Employment, Native Non-College-Educated**	**(5)** **K-P Wald *F*-Statistic of the First Stage**
Basic: O*NET S&E state and year effects	4.97*** (1.15)	–0.15 (0.91	2.04 (1.54)	–3.04 (2.66)	17.86
Major-based S&E, metro-area and year effects	6.84*** (1.93)	1.27 (1.41)	0.69 (1.51)	–5.92 (3.27)	20.02
As Basic, omitting San Jose, Calif., and Stamford, Conn.	6.31*** (1.73)	–1.03 (1.59)	1.51 (1.48)	–5.13 (4.57)	42.07
As Basic, omitting NY, LA, SF, and DC	4.58*** (1.16)	0.2 (0.87)	2.38 (1.73)	–2.04 (2.75)	15.18
As Basic, using decade changes	4.20*** (1.26)	–1.02 (0.97)	1.76 (1.22)	–2.45 (2.54)	20.28
As Basic, adding control for 1980 share of college-educated	3.97** (1.47)	–0.63 (1.28)	–1.16 (1.86)	–5.45 (3.65)	9.25
Imputation of IV-based on 1970 foreign- major-based S&E	5.20*** (1.37)	–1.24 (1.43)	–2.92 (1.87)	–8.33 (4.46)	6.05
Total S&E as explanatory variable	3.99*** (1.64)	–0.12 (0.75)	1.64** (0.75)	–2.44 (3.08)	4.34
OLS, basic specification	2.50*** (0.83)	1.17*** (0.41)	3.52** (1.53)	3.19** (1.22)	N/A

Note: Each cell includes the estimate of the effect of growth of foreign-S&E (or total S&E) on the dependent variable listed at the top of the column. The instrument used is the H-1B-driven growth of foreign S&E workers. The basic specification estimated is described in the text. It always includes state and period effects and, where not absorbed by the fixed effects, it also controls for the 1980 employment share of native S&E. The specification using 1970 IV does not include the state fixed effect. The last row shows the OLS estimate of the basic specification. The standard errors are heteroskedasticity-robust and clustered at the metro-area level. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively. DC = Washington, D.C.; IV = instrumental variable; LA = Los Angeles; N/A = not applicable; NY = New York; OLS = ordinary least squares; O*NET = Occupational Information Network; SF = San Francisco.

The different rows of Table 6 represent different specifications and samples. Each includes period effects, state or metro area effects, and—when not absorbed by the metro-area effect—the 1980 native S&E share of employment. In this section, we focus on the first row, which reports 2SLS estimates from the basic specification using the O*NET S&E-based definition of occupations. In this specification, we use the 1980-based H-1B variable as an instrument for the actual foreign S&E growth.

Four interesting results emerge, and they are then confirmed by robustness checks. First, there is a positive and significant effect of H-1B-driven increases in S&E on wages paid to college-educated workers. The estimated effect is significantly different from zero at the 1 percent level, and the point estimate is close to 5, implying a 5 percentage point increase in the growth rate of wages for each increase in the H-1B-driven S&E growth by 1 percentage point of initial employment.[28] This elasticity seems large; however, recall that foreign S&E are a very small fraction of employment, and they grew by only half a percent of employment over the 20 years from 1990 to 2010. Second, H-1B foreign workers did not have any significant effect on wages of non-college-educated workers. The point estimate is much smaller than that for college-educated wages and is insignificantly different from zero. Third, the inflow of foreign S&E did not significantly affect the employment of college-educated natives. Fourth, H-1B-driven S&E growth had no significant effect on non-college-educated employment. The point estimate in column (4) (row 1) is negative, but it is also imprecisely estimated and not significant. Importantly, the remaining rows of Table 6 that use 2SLS estimation (discussed in greater detail in the next section) find the same four regularities, pointing to the robustness of the results.

The null effect on non-college-educated workers and the positive wage effect on college-educated workers together suggest that H-1B workers contributed to skill-biased productivity growth. With regard to our simple model of Figure 4, those changes reveal a positive

[28] In nonreported regressions, we also estimated the effect of foreign S&E on wages of native S&E only. The results are, in most cases, similar to the effect on native college-educated workers. Hence, we considered college-educated natives (S&E and non-S&E) as a homogeneous group.

shift to the labor demand for native college-educated workers and no shift of the demand for non-college-educated workers. The weak employment response of college-educated natives might also suggest that other adjustment mechanisms were at work at the metropolitan-area level, thus implying a rather rigid labor supply. In a related check (not reported), we find that the rental cost of housing for college-educated workers was positively affected by the inflow of foreign S&E in the metropolitan areas.[29] A new spatial equilibrium in which local prices (rents) respond to wages would imply little labor mobility. Ottaviano and Peri (2006) find a similar effect of immigrant diversity on wages and rents.

It is particularly important to understand the magnitude of the one significant S&E coefficient in row 1 of Table 6: the 4.97 value in the college-educated natives' wage regression. We offer several alternatives. First, we can combine the results of Tables 5 and 6 to better understand the role of the H-1B visa. A shock of 1 standard deviation to our imputed H-1B growth variable across cities and time periods equals 0.00038. From column (1) of Table 5, we see that shock would translate into a 0.000266 (= 0.00038 × 0.70) increase in foreign S&E. That would then imply a 0.00132 (= 0.000266 × 4.97) rise in the growth rate of real wages paid to college-educated natives. That figure amounts to roughly 3.4 percent of a standard deviation of the dependent variable, a seemingly small response.

Interpretation at the national level, in contrast, suggests a larger wage response. First, note that foreign S&E represent only 1–3 percent of employment, depending on how they are measured. Moreover, their growth was only about 0.4 percentage points of total employment during the 1990s and 0.3 percentage points in the following decade. Applying the 2SLS estimates of row 1 in Table 6 to the average growth in foreign S&E nationally implies that the foreign-driven net increase in S&E increased real wages of college-educated natives by about 2 percentage points (0.4 × 4.97) between 1990 and 2000,

[29] A regression of the median rental cost per room on the policy-instrumented increase in foreign S&E shows that a 1 percentage point increase in the policy instrument increased rents for college-educated individuals by 4.11 percentage points (standard error 1.29). For non-college-educated individuals, the effect was a nonsignificant −1.90 percentage points (standard error 1.45). Hence, part of the purchasing power from increased wages for the college-educated was absorbed by increased rental costs.

and by an additional 1.5 percentage point (0.3 × 4.97) between 2000 and 2010. Thus, overall foreign S&E increased the cumulated wage growth of college-educated natives by 3.5 percentage points between 1990 and 2010. From our census data, we can calculate that the average real weekly wage of U.S. college-educated workers increased by a cumulative 13 percentage points over that 20-year period. The wage of non-college-educated workers, in contrast, remained essentially stagnant. Therefore, about one-fourth of the college-educated wage growth can be attributed to productivity growth from foreign S&E. Given that foreign labor accounted for nearly a quarter of U.S. science and engineering employment in 2010 (see Table 2) and that science and engineering are primary drivers of U.S. productivity and wage growth, we find those estimates quite plausible. In Section 7, we will compare the foreign S&E-related wage gains to those related to other growth determinants.

Before we discuss the other rows of Table 6 in the next section, it is worth highlighting that a comparison between the first and last rows of Table 6 provides insight into the validity of our causal estimates. The last row shows the OLS analogue of the 2SLS results in the first row; coefficients obtained by regressing the dependent variables on foreign S&E growth, state effects, and period effects. The OLS regression finds positive and significant effects of S&E on all variables with higher values (compared with 2SLS results) in three of the four cases. S&E are attracted to cities in which employment and wages of all workers are growing (reverse causality).

Our instrument, in contrast, allows us to separate S&E's positive effect on the demand for college-educated workers from the null effect on the demand for non-college-educated labor. Our intuition is that the endogeneity bias—namely, the tendency to have positive effects in boom cities—may be very strong for employment and for less educated workers' wages, therefore contributing to a more severe upward bias in OLS estimates. Somewhat surprising, however, the OLS estimate of S&E's effect on wages paid to college-educated natives is actually smaller than the 2SLS estimate. In this case, we suspect that endogeneity bias itself might be comparatively small. Measurement error, especially in the construction of the relevant S&E share from an imperfect collection of occupations—common to all variables—could be responsible for the downward bias visible in the college-educated native wage regressions.

5.2 Robustness Checks

The remaining rows of Table 6 show the estimates of the same four coefficients ($b_{w,Coll}$, $b_{w,no\text{-}Coll}$, $b_{E,Coll}$, and $b_{E,no\text{-}Coll}$) using different specifications. In row 2, we use the S&E definition that is based on college major, and, given that the instrument is quite powerful (as noted in Table 5), we also include the metropolitan-area fixed effects. In those two specifications, we also modify the H-1B-based instrument accordingly. The *F*-test of this specification is stronger than for the basic one (row 1).

The rows 3 and 4 explore the sensitivity of the regression to outliers. First, we omit metropolitan areas with unusually large growth in foreign S&E (row 3). Then, we omit metro areas with very large sizes and hence large numbers of foreign S&E relative to the national total. The results are qualitatively and quantitatively very similar to the basic specification. Though we prefer to think of 1990–2000, 2000–2005, and 2005–2010 as three distinct periods (in light of the very different economic conditions introduced in the second half of the 2000s by the Great Recession), we also recognize that it is somewhat unusual to use three panels of nonuniform length.[30] Row 5 therefore uses a panel of two decade-long changes (1990–2000 and 2000–2010) instead.

Row 6 adds a control for the 1980 share of college-educated people in the population. The control captures the human capital intensity of the city that has been found to be important in determining its growth and can be correlated with the foreign S&E share. Row 7 uses the 1970-based instrument that, as illustrated in Table 5, might suffer from weak-instrument problems. Moreover, to have sufficient power in this case, we do not include state fixed effects. The standard errors, especially in column (4), are quite large. Still, the main results are very consistent with those obtained in the basic specifications. Finally, row 8 shows the result when we consider the total change in S&E (rather than foreign only), which is still instrumented with the H-1B policy measure.

The main results and the point estimates of row 1 are clearly confirmed in the robustness checks. Only the effect on wages paid to college-educated natives is significantly different from zero in each

[30] Borjas, Grogger, and Hanson (2012) and Ottaviano and Peri (2012) also use a panel that includes periods of different lengths because of data availability.

specification. Estimates for that effect range between 3.9 and 6.8. Let us emphasize that several of our specifications are very demanding by including a full set of 219 metropolitan areas' fixed effects in a differenced panel (with only three periods). The coefficients are identified on differences in the growth rates of S&E in a city across periods. The main characteristics of the coefficients in all robustness checks are consistent with those of the basic specifications. Finally, the estimated effects on non-college-educated employment are negative but not significant in any specification. The standard errors of those estimates, however, are usually large.

5.3 Extensions

Taking all the specifications together, our estimates reveal that the demand for native college-educated workers received a significantly positive boost from foreign S&E. At the same time, however, the demand for non-college-educated labor was not positively affected. In this section, we analyze in greater detail how S&E might have affected the city's economy beyond the broad groups previously considered. First, we analyze whether the null effect on the demand for non-college-educated labor is concentrated mainly in the very low part or the intermediate part of the educational distribution. That is, we assess whether the wage and employment effects are different among individuals with and without a high school diploma, or if instead the effects are uniform across all non-college-educated natives. Second, we analyze whether S&E growth pushed college-educated native workers toward specific sectors.

Table 7 shows the effect of foreign S&E on the weekly wages (columns [1] and [2]) and employment (columns [3] and [4]) of native workers without a college education. We separate between high school dropouts (columns [2] and [4]) and high school graduates (columns [1] and [3]). In rows 1 and 2, we show the estimates from the 2SLS regression—either in its basic specification or using the growth of total S&E as the explanatory variable—instrumented with the H-1B policy variable. The last row reports the OLS coefficients. By distinguishing high school graduates from high school dropouts, we can check whether the two groups are differentiated in their responses to productivity shocks brought on by foreign S&E. It is also a test for whether S&E workers produced the type of change in labor demand often referred to as the polarization of the labor market.

Table 7

Effect of Foreign-S&E on Wages and Employment of Native Workers: Split Non-College-Educated into Two Groups (219 U.S. metropolitan areas; 1990–2000, 2000–2005, and 2005–2010)

	Dependent Variable: Growth Rate of				
Explanatory Variable: Growth rate of foreign (total) S&E **Instrument:** H-1B Imputed Growth of Foreign-S&E	**(1)** **Weekly Wage, Native HS Graduates**	**(2)** **Weekly Wage, HS Dropouts**	**(3)** **Employment, Native HS Graduates**	**(4)** **Employment, Native HS Dropouts**	**(5)** **K-P Wald *F*-Statistic of the First Stage**
Basic: O*NET S&E state and year effects	0.25 (0.96)	−3.84 (2.41)	−3.57 (2.35)	0.53 (0.46)	17.86
Total S&E as explanatory variable	0.2 (0.73)	−3.09 (2.48)	−2.87 (2.98)	0.43 (0.27)	4.34
OLS, basic specification	1.21** (0.44)	0.30 (1.30)	2.51** (1.07)	0.69** (0.20)	N/A

Note: Each cell includes the 2SLS estimate of the impact of growth of foreign S&E (or total S&E) on the dependent variable listed at the top of the column. The instrument used is the H-1B-driven growth of foreign S&E. Each specification includes state and period effects and controls for the 1980 share of native S&E. The last row shows the OLS estimate of the basic specification. The standard errors are heteroskedasticity-robust and clustered at the metro-area level. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively. HS = high school; N/A = not applicable; OLS = ordinary least squares; O*NET = Occupational Information Network.

This phenomenon is characterized by higher employment growth at the high and low ends of the education spectrum at the expense of intermediate-level jobs (see, e.g., Autor 2010; Autor, Katz, and Kearney 2006). The estimates of Table 7 show that S&E effects on both high school graduates and dropouts are mostly negative and insignificant. Those estimates, therefore, do not support local positive employment effects for the least educated. They are instead consistent with claims that science and engineering–driven technological progress contributed to the skill bias of the labor market (rather than polarization) in the United States.

Table 8 shows how employment of native college-educated workers (column [2]) and all native workers (column [1]) across nine separate sectors responded to foreign S&E. We include all sectors except those that have very small employment shares in some cities and would, therefore, exhibit rather noisy estimates (e.g., mining, agriculture, and entertainment). We arrange sectors in Table 8 in three groups: (a) private sectors with low human capital intensity (measured as having a share of college-educated labor smaller than 25 percent in the year 2000), (b) private sectors with high human capital intensity (measured as having a college-educated share larger than 25 percent), and (c) the public sector (whose employment growth may not be driven by productivity considerations). The coefficients of column (2) in Table 8 are obtained using the basic specification (as in row 1 of Table 6), and they show that the employment of college-educated labor in high human capital sectors increased much more than in low human capital sectors in response to an inflow of S&E into the city. The college employment response is significant in two out of three sectors. In contrast, low human capital sectors and the public sector did not experience any significant college-educated job growth. Hence, cities with high S&E inflows experienced a reallocation of college-educated workers toward more human capital–intensive sectors. The coefficients of column (1) show that although two of the high human capital sectors experienced positive total employment changes (i.e., employment increases among workers of all education levels) in response to S&E, those effects were not significant.

In summary, our empirical analysis identifies a positive, large, and significant effect of foreign S&E growth on the growth of college-educated native wages across U.S. metropolitan areas. At the

Table 8
Effects of Foreign S&E on Employment by Industry (219 U.S. metropolitan areas; 1990–2000, 2000–2005, and 2005–2010)

	(1)	(2)
Explanatory Variable: Growth rate of foreign S&E	**Dependent Variable: Total Employment**	**Dependent Variable: College–Educated Employment**
Instrument: H-1B imputed growth of foreign S&E	**2SLS**	**2SLS**
Low Human Capital, Private Sector		
Construction	−0.23	0.02
	(0.25)	(0.03)
Transportation	−0.19	0.05
	(0.27)	(0.03)
Wholesale	−0.11	−0.04
	(0.11)	(0.03)
Manufacturing	0.46	0.06
	(0.76)	(0.17)
Retail	−0.14	0.1
	(0.48)	(0.07)
High Human Capital, Private Sector		
Finance	0.14	0.2
	(0.24)	(0.15)
Business	0.13	0.38***
	(0.15)	(0.08)
Professional Services	−0.28	0.68*
	(0.64)	(0.36
Public Sector		
	−0.25	−0.01
	(0.20)	(0.06)

Note: Each cell includes the 2SLS estimate of the effect of the growth of foreign S&E on the dependent variable listed at the top of the column within the sector listed in the row. Each regression includes state and year fixed effects. The instrument used is the H-1B-driven growth of foreign S&E workers (using the major-based definition of S&E). The standard errors are heteroskedasticity-robust and clustered at the metro-area level. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

same time, we do not find any significant effect on the growth of wages and employment of non-college-educated natives. The employment of college-educated natives did not exhibit overall growth in response to foreign S&E, but instead shifted from sectors with low human capital content to high human capital sectors.

6. THE EFFECTS OF FOREIGN S&E IN CANADIAN METROPOLITAN AREAS

The estimates presented earlier show an important productivity effect of foreign S&E on the U.S. economy. The United States, however, is at the technological frontier and is the largest international magnet for high-quality S&E. In this section, we assess the Canadian experience as an alternative in order to test whether other countries experience similar productivity boosts from attracting international S&E. Although Canadian cities and provinces do not have the same number and concentration of innovation-based firms and institutions as the United States, Canada's FSWP has been structured to attract highly educated workers with a strong weight given to S&E. Hence, we use the same approach used for the United States, and we construct from Canadian census data the imputed inflow of foreign S&E in each of 17 geographical areas (mostly metropolitan areas) between 1990 and 2006, according to the distribution of foreign S&E by nationality in 1980 and the aggregate inflow of foreign S&E by nationality. Then, we use that imputed immigration factor as an instrument for the actual increase in the number of foreign S&E as a share of employment across geographical areas.

The Appendix describes in detail the Canadian data and the results from the first-stage regressions. Here, we simply note that despite the small number of observations (17) over three five-year periods (1991–1996, 1996–2001, and 2001–2006), the imputed-immigrant instrument has reasonably strong power in predicting foreign S&E growth in Canadian metropolitan areas. The F-statistic is equal to 14.05 for the basic specification, which indicates that the power of the Canadian instrument is comparable to that for the U.S. data.

Table 9 shows the estimated effects of foreign S&E in Canadian geographical areas for three different specifications. The structure of Table 9 is the same as that of Table 6. Each specification includes period fixed effects and controls for the 1980 employment share of native S&E. Row 1 reports results from the basic specification. The

Table 9

2SLS Regression of Foreign S&E on Wages and Employment of Native Workers: Canada (17 Canadian geographic areas; 1991–1996, 1996–2001, 2001–2006)

Explanatory Variable: Imputed growth in foreign S&E	**(1)** **Growth Rate in Weekly Wage, Native College-Educated**	**(2)** **Growth Rate in Weekly Wage, Native Non-College-Educated**	**(3)** **Growth Rate in Employment, Native College-Educated**	**(4)** **Growth Rate in Employment, Native Non-College-Educated**	**K-P Wald *F*-Statistic of the First Stage**
Basic specification, with period effects	5.55** (2.67)	2.96 (1.86)	0.53* (0.27)	0.39 (1.63)	14.05
Basic specification, with province fixed effects	2.80** (1.19)	0.54 (0.85)	0.80** (0.32)	−1.22 (0.86)	4.79
Total change in S&E as explanatory variable	4.89** (1.82)	2.61** (1.28)	0.47** (0.13)	0.35 (1.36)	6.43

Note: Each cell includes the 2SLS estimate of the effect of growth of foreign (or total) S&E on the dependent variable listed at the top of the column. The instrument used is the imputed growth of foreign S&E. The basic specification estimated includes period effects and the 1980 native S&E as share of employment. The standard errors are heteroskedasticity-robust and clustered at the province level. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

specification in row 2 is the same as in row 1 but adds province effects. Row 3 uses the total change in S&E (native and foreign) as an explanatory variable, still instrumented by the imputed inflow of foreign S&E.

The estimated coefficients are qualitatively and quantitatively close to those estimated for the United States in Table 6. First, there is a positive and significant effect of foreign S&E on wages paid to Canadian college-educated workers. The estimated effect is always statistically different from zero at the 5 percent significance level, and the point estimates are between 2.8 and 5.5 percentage points for each percentage-point increase in the S&E share of employment.

The second consistent result is that foreign S&E also had a positive effect—although not always significant at the 5 percent level—on the employment of native college-educated workers. The point estimates of that effect are smaller than those on wages, but they are consistent with the fact that an increase in productivity of highly educated workers in cities experiencing large inflows of foreign S&E might have attracted more college-educated natives in those cities. It is possible that local prices in Canada did not respond as much as in the United States, and hence the inflow of highly educated workers was larger.

The third important takeaway is that the inflow of foreign S&E generally had an insignificant effect on the wages and employment of non-college-educated natives. The insignificant effect on non-college-educated workers and the positive wage effect on college-educated natives and on their employment, together, suggest that foreign S&E also had a positive productivity effect in Canadian geographical areas, with a skill bias. It is, however, worth looking more carefully at the effects within non-college-educated labor to see whether some subgroups benefited while others lost. Table 10 shows the effect of foreign S&E on the wages (columns [1] and [2]) and employment (columns [3] and [4]) of native non-college-educated workers. In the specifications, as in Table 7 for the United States, we refine this educational category by separating workers into high school dropouts (columns [2] and [4]) and high school graduates (columns [1] and [3]). That distinction allows us to check whether the two groups are differentiated in their complementarity with college-educated labor.

The estimates of Table 10 show that the S&E effects on high school graduates are insignificant (the point estimate is even negative

Table 10
Foreign S&E Effect on Wages and Employment: Split Non-College-Educated into Two Groups, Canada (17 Canadian geographic areas; 1991–1996, 1996–2001, 2001–2006)

	(1)	(2)	(3)	(4)
Specification:	**Weekly Wage, Native High School Graduates**	**Weekly Wage, Native No High School Diploma**	**Employment, Native High School Graduates**	**Employment, Native No High School Diploma**
Imputed change in foreign S&E	2.15 (1.57)	4.28 (2.70)	−0.49 (1.40)	0.89** (0.27)
Observations	51	51	51	51
Period effects	Yes	Yes	Yes	Yes
K-P Wald *F*-Statistic of the first stage	14.05	14.05	14.05	14.05

Note: Each cell includes the 2SLS estimate of the effect of growth of foreign S&E on the dependent variable listed at the top of the column. The instrument used is the imputed growth of foreign S&E workers. The specification estimated includes period effects. The standard errors are heteroskedasticity-robust and clustered at the province level. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

for the employment effect). Conversely, the effects on wages and employment of high school dropouts are positive, and the first is close to 10 percent significance, while the second is significant at 5 percent. In particular, a 1 percentage point increase in the growth of the foreign S&E share of employment raises wage growth of native high school dropouts by 4.2 percentage points and their employment growth by 0.9 percentage points of total employment. This result is consistent with claims that S&E-driven technological progress contributed to labor-market polarization by positively affecting the demand for low-education groups (dropouts) more than for intermediate-education groups (high school graduates).

That positive and significant effect on employment of workers without a degree is an interesting sign that the benefits from S&E may diffuse in the local economy, via local demand, to workers not

directly affected by the increases in productivity driven by information technology. Foreign scientists and engineers benefit college graduates, but they possibly benefit workers with low skill levels through local labor demand linkages as well.

That effect on the productivity of less educated Canadian labor was not found in the United States. That finding suggests that foreign S&E may have contributed to labor-market polarization in Canada. Technological innovation spurred by S&E may have substituted for routine cognitive tasks typically performed by workers with a high school diploma (e.g., data processing and logistical, organizational, and secretarial activities). In contrast, however, this type of innovation did not decrease the demand for manual tasks in sectors such as construction, personal care, hospitality, and home services that are typically performed by workers who are not high school educated. Hence, IT innovation did not displace the lowest-educated workers and, in fact, may have even created new jobs for those workers as college-educated individuals became richer and demanded more services. Understanding why foreign S&E would affect the least-educated workers in the United States and Canada differently is an interesting topic for future research.

7. COMPARISON WITH OTHER LOCAL GROWTH POLICIES

Attracting S&E is just one of the possible strategies often considered by local governments to promote productivity growth for a local economy. In fact, U.S. federal and state spending on such place-based policies amount to $60 billion each year (Moretti 2010). Urban economists have recently emphasized that the presence of innovative, fast-growing, and highly productive plants and industries (Greenstone, Hornbeck, and Moretti 2010; Moretti 2010) and the presence of college-educated workers (Iranzo and Peri 2009; Moretti 2004a) as two important strategies that promote local productivity growth. In this section, we control for proxies of those two growth-promoting factors and compare their effects with those from increases in H-1B-driven S&E.

Let us also notice, before proceeding further, that our identification of the productivity effect of foreign S&E has relied on measures that encompass only effects within the metropolitan areas. If S&E had productivity spillovers that go beyond the boundaries of the metropolitan area, we would miss those effects. Hence, if spillovers

are more global in nature, our estimates miss a significant part of the effects, and they represent lower-bound estimates of the national productivity effect of S&E.

7.1 The Effect of High-Growth Industries

As a growth strategy, urban planners have long pursued local agglomeration of human capital–intensive and fast-growing industries. For example, Silicon Valley is often cited as a successful and productivity-enhancing agglomeration of the computer and IT sectors, and several locations have tried to reproduce its model. Although it is not clear what ingredients are needed to create such a productive agglomeration, most economists consider the presence of companies in dynamic and innovation-intensive sectors and the availability of college-educated workers as key factors for productivity growth.

According to that view, we construct a control variable for our regressions to account for sector composition. First, we take the sector composition of each metro area (indexed by c) in 1980 as given, and record the share of employment in each of 243 sectors (indexed by i) spanning manufacturing and services, $s_{i,c,1980}$. Then, we construct the growth of average wages and employment that would have occurred if each industry grew at its industry-specific national average rate between 1980 and 2010. We call those variables the "Bartik" sector-driven wage and employment growth (from Bartik 1991) in the metropolitan area.

The variables are mathematically described in equations (7) and (8):

$$\left(\frac{\Delta w_{c,t}{}^{Bartik}}{w_{c,t}{}^{Bartik}}\right) = \frac{\sum_{i=1,243} s_{i,c,1980}\, w_{i,t}}{\sum_{i=1,243} s_{i,c,1980}\, w_{i,t-1}} - 1 \tag{7}$$

$$\left(\frac{\Delta Emp_{c,t}{}^{Bartik}}{Emp_{c,t}{}^{Bartik}}\right) = \frac{\sum_{i=1,243} s_{i,c,1980}\, Emp_{i,t}}{\sum_{i=1,243} s_{i,c,1980}\, Emp_{i,t-1}} - 1 \tag{8}$$

In those two equations, $s_{i,c,1980}$ is the share of workers in city c who are employed in industry t as of 1980.[31] The terms $w_{i,t}$ and $Emp_{i,t}$ are national average wages and national total employment of workers in industry i in year t. The effect of those variables on the metropolitan area's employment and wage growth captures the effect from the

[31] Formally, $s_{i,c,1980} = \frac{Emp_{i,c,1980}}{Emp_{c,1980}}$.

1980 industrial structure of metropolitan areas. The advantage is that we can think of the predetermined 1980 industrial structure as the cause for higher or lower growth, depending on the larger or smaller share of sectors that have subsequently grown at high rates.

Evaluating the effect of those constructed variables on wages and employment allows us to establish how important the industrial structure of a metro area in 1980 was for the subsequent growth of its employment and wages. We also explore the extent to which attracting fast-growing industries, such as the computer industry, would have changed wage and employment growth in the city. At the same time, we can compare the effect of that sector-composition variable on wage and employment growth with the effect of the H-1B-driven growth in S&E. To do so, we estimate a direct regression in which we include the H-1B policy variable as well as the sector-driven employment and wage growth. Namely, we estimate

$$(9) \quad y_{c,t}^{Native,X} = \varphi_t + \varphi_s + b_{y,X} \frac{\Delta(S\&E)_{c,t}^{H\text{-}1B}}{Emp_{c,t}} + d_1^{y,X} \frac{\Delta w_{c,t}^{Bartik}}{w_{c,t}^{Bartik}} + d_2^{y,X} \frac{\Delta Emp_{c,t}^{Bartik}}{Emp_{c,t}^{Bartik}} + \gamma \cdot Controls_{c,t}^{X} + \varepsilon_{c,t}$$

In equation (9), the term $y_{c,t}^{Native,X}$ represents alternatively the wage or employment growth for group X (college-educated or non-college-educated). We directly include as explanatory variables the H-1B-driven change in foreign S&E as well as the wage and employment Bartik instruments. The remaining terms in equation (9) are the same as in equations (1) and (2), which include period (φ_t) and state effects (φ_s) and a control for the 1980 native S&E share of employment. We estimate this specification by least squares and cluster standard errors at the metro-area level.

The upper part of Table 11 shows the coefficients of the three explanatory variables on the growth of weekly wages of native college-educated workers (column [1]), on the growth of weekly wages of native non-college-educated workers (column [2]), on the employment growth of college-educated workers as a share of total employment (column [3]), and on the employment growth of non-college-educated workers as a share of total employment (column [4]). The first finding that emerges is that the effects of the H-1B-driven policy variable are not altered when we control for the Bartik instrument. The H-1B variable retains a very significant effect on the wage of college-educated workers and an insignificant effect on the wages and employment of non-college-educated labor.

Table 11

Reduced-Form Regressions of Growth Determinants: H-1B Policy, Industrial Composition, and College Enrollment Rate
219 U.S. metropolitan areas; 1990–2000, 2000–2005, and 2005–2010

Dependent Variable:	(1) Growth Rate in Weekly Wage, Native College-Educated	(2) Growth Rate in Weekly Wage, Native Non-College-Educated	(3) Growth Rate in Employment, Native College-Educated	(4) Growth Rate in Employment, Native Non-College-Educated
H-1B Policy and Sector Composition				
H-1B-driven growth of foreign S&E	3.42***	−0.21	1.51	−2.05
	(0.88)	(0.64)	(1.26)	(1.84)
Industry-predicted employment growth	−0.09	0.25	0.35	1.16***
	(0.50)	(0.25)	(0.25)	(0.49)
Industry-predicted wage growth	0.09	0.35***	0	0.43
	(0.34)	(0.13)	(0.25)	(0.26)
H-1B Policy and Past College Enrollment Rates				
H-1B-driven growth of foreign S&E	3.37***	−0.17	1.27	-2.58
	(0.89)	(0.64)	(1.18)	(1.86)
Lagged college enrollment rate	0.05	0.03	0.07	0.20*
	(0.07)	(0.04)	(0.04)	(0.10)

Note: The upper part of the table shows the estimates from regressions, including the H-1B policy variable and industry-predicted growth as explanatory variables. The lower part shows the estimates from regressions with the H-1B policy variable and the five-year lagged enrollment rate in college as an explanatory variable. Each regression includes year effects, state effects, and native S&E as a share of 1980 employment. The standard errors are heteroskedasticity-robust and clustered at the metro area level. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

Second, the presence of industries that experienced fast employment growth nationally contributed to the employment growth of metro areas, especially for non-college-educated workers. For each 1 percentage point of faster sector-driven employment growth, a metro area experienced faster college-educated employment growth by 0.35 percentage points (not significant) and non-college-educated employment growth by 1.16 percentage points (significant at the 1 percent level).

Similarly, the presence of industries with fast national wage growth contributed to city-level wage growth for non-college-educated workers (by 0.35 percentage points for each percentage point of sector wage growth). Those results imply that the industrial structure of a city—in particular, the presence of fast-growing industries—was more relevant than the inflow of H-1B S&E in boosting wages and employment of non-college-educated labor. However, the H-1B policy variable seems to have the largest and most significantly positive effect on wages paid to college-educated workers. Thus, attracting S&E seems to be a key to productivity growth for the college-educated.

As the units for the explanatory variables are different, we cannot directly compare the magnitude of the estimated coefficients in Table 11. Therefore, we provide some calculations of the magnitude of each effect. The standard deviation across metro areas for both the industry-imputed growth of employment and of wages (Bartik) in a decade was about 0.01. The standard deviation for the H-1B imputed policy variable across metro areas in one decade was 0.004. Hence, an increase in the sector-driven wage variable by 1 standard deviation produced an increase in wage growth of non-college-educated workers of 0.35 percentage points in a decade, while an increase in the H-1B imputed inflow of S&E increased the wage growth of college-educated workers by 1.4 ($= 0.004 \times 3.42$) percentage points in the decade. A similar increase of 1 standard deviation in the employment-imputed Bartik would also increase employment growth of non-college-educated labor by 1.16 percentage points in a decade. Those effects are economically significant.

Perhaps a more interesting and policy-relevant exercise is to consider a specific case, such as the effect of increasing the size of the computer sector[32] in a metro area from its average value of 1.5 percent

[32] We consider the computer sector to be two industries called "computers and relative equipment" and "computer and data-processing services."

of total employment in 1990 to the value in San Jose, Calif.—home to Silicon Valley—of 10 percent of employment. Because the computer sector experienced much faster than average employment and wage growth, this jump from the national average to the Silicon Valley level would be captured by increases in our predicted (Bartik) employment and wage growth by 4 percent and 1 percent per decade, respectively. Those two changes together imply (using the statistically significant coefficients in Table 11) a 1.4 (= 4 × 0.35) percentage point faster growth for non-college wages during the decade and a 1.16 percentage point faster growth for non-college employment because of the larger computer industry (0.43 × 4 + 1.16). There would be, however, no significant effect on wages and employment of college-educated labor.

In comparison, if we were to change the H-1B-driven growth of foreign S&E in a city from the average value in the 1990–2000 decade (0.4 percent) to the value in San Jose (2 percent), the coefficient in Table 11 suggests that change would result in an increase in the wage growth of college-educated workers by 5.47 (= 1.6 × 3.42) percentage points in the decade. Overall, the inflow of foreign S&E had the most significant effect in raising the wage growth of college-educated native workers in U.S. cities. The presence of fast-growing industries, instead, seems to have affected employment growth in a broader way, especially for the non-college-educated, as well as for wages paid to non-college-educated workers.

7.2 The Effect of College Enrollment

College-educated workers are an important determinant of local wages and productivity through local externalities.[33] That is the reason we introduced the 1980 share of college-educated workers in one specification of Table 6: so we could control for preexisting human capital agglomeration and its effect on productivity growth. In this section, we consider city-level policies designed to attract workers with a college degree. Although explicitly targeting the inflow of college-educated workers is a rare policy, many cities attempt to increase the college enrollment of its college-aged citizens—either by increasing the number of local colleges or by expanding their size—in hopes that those individuals will graduate and improve the local work force. Therefore, we test enrollment rates lagged by

[33] See, for instance, Moretti (2004a).

five years as a predictor of a city's wage and employment growth. We do recognize that policies aimed at increasing enrollment will likely produce smaller effects than expected if the mobility of recent college graduates were not accounted for, since college students are not mandated to work and live in the same city after graduation.[34]

The bottom part of Table 11 shows the effect of the H-1B policy variable together with the effect of college enrollment rates, calculated as the number of individuals enrolled in college between ages 18 and 25 relative to the working-age population (18–65) in the period (10 or 5 years) before the considered interval. The regressions include the usual 219 metro areas in the time periods 1990–2000, 2000–2005, and 2005–2010. The estimated coefficients in rows 4 and 5 of Table 11 show that the inclusion of lagged college enrollment rates does not change the effect of H-1B visa-driven S&E. Enrollment itself has only a very mild effect on the employment growth of non-college-educated workers in the metro area. An increase in the lagged college enrollment rate by 1 standard deviation (0.05) would increase employment growth by 1 percentage point for non-college-educated labor. The effect on metro area wages is negligible.

7.3 Combining Policies

It is interesting to ask whether the effectiveness of the H-1B policy on native wages is particularly strong in metropolitan areas with the "right" sector composition. After all, foreign-born S&E have fueled the growth of the information, communication, and computer sectors, and it makes sense to check whether the productivity effects were stronger in metro areas where those sectors were large. That exercise will speak to the opportunity of potentially combining policies for growth (attracting industries as well as promoting immigration of S&E). To check the productivity effects, we construct a dummy that equals 1 in metro areas that experienced fast sector-driven employment growth. In particular, if a metro area had a sector-predicted (Bartik) employment growth of 20 percent or more in a decade, the

[34] Lagged enrollment also helps avoid spurious correlation caused if unobserved factors affect productivity and current college enrollment. Lagged enrollment is somewhat predetermined with respect to demand shocks taking place five years later. Moretti (2004a) uses the establishment of land-grant colleges in a metropolitan area as the exogenous factor affecting the share of college-educated workers decades later.

dummy is 1; otherwise it is 0. The dummy captures those metro areas with a large share of employment in sectors such as "computers and relative equipment" that grew very quickly during the observed time period. For about 25 percent of the observations, the dummy equals 1. We then interact that dummy with the H-1B-driven growth of S&E and include that interaction in a regression otherwise similar to Table 11.

Column (1) of Table 12 shows the estimated effects on the real wage growth of native college-educated labor when including that interaction. Very interestingly, we notice that the whole productive effect of the H-1B policy is concentrated in the cities with rapidly growing industries. The main effect of the H-1B variable is not statistically significant, whereas an increase of H-1B among the cities with fast-growing sectors (interaction term) by 1 percentage point of employment increased wages of college-educated natives by 8.3 percentage points. The combination of a high inflow of H-1B labor and the presence of fast-growing sectors is a powerful combination for wage growth.

Column (2) of Table 12 shows results from a regression interacting H-1B growth with an indicator for cities exhibiting a large lagged share of college enrollment (larger than 0.12), otherwise similar to the one in the lower part of Table 11, column (1). The estimated coefficients do not reveal any stronger effect of foreign S&E in cities with larger college enrollment.

7.4 Simple Policy Calculations: The Long-Run Effect of Raising the H-1B Cap

We close our analysis by combining the estimated effects of Table 11 and the average inflow of H-1B visas between 1990 and 2010 to calculate the overall effect of that program on the wages of native college-educated workers. First, the average estimates from Table 11 suggest an effect of H-1B-driven S&E on the wage of native college-educated labor equal to 3.4. Second, the average yearly inflow of H-1B visas during the same two decades (about 97,000 visas per year on average) corresponds to an annual increase of 0.05 percentage points of the labor force. Together, they suggest that a continued H-1B-induced inflow of S&E produced a faster growth of wages paid to native-born college-educated labor by ($0.05 \times 3.4 =$) 0.17 percentage points per year. Over 20 years, that result translates to real wage

Table 12
Interactions between H-1B Policy and Industrial Composition or College Enrollment Rate (219 U.S. metropolitan areas; 1990–2000, 2000–2005, and 2005–2010)

Dependent Variable: Growth rate in weekly wage, college-educated natives	**(1)** **Interaction H-1B-Sector Composition**	**(2)** **Interaction H-1B-College Enrollment**
H-1B Policy and Sector Composition		
H-1B-driven growth	−0.90	
	(0.99)	
Industry-predicted employment growth	−0.15	
	(0.50)	
H-1B-driven growth × high predicted employment growth	8.30***	
	(2.16)	
H-1B Policy and Past College Enrollment Rates		
H-1B-driven growth		3.37***
		(0.91)
Lagged college enrollment rate		0.05
		(0.09)
H-1B-driven growth × high predicted employment growth		0.01
		(2.27)

Note: The dependent variable is the growth rate in weekly wage of college-educated natives. Each specification includes year effects, state effects, and the native S&E share of 1980 employment. The standard errors are heteroskedasticity-robust and clustered at the province level. ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

gains equal to 3.5 percentage points. If we consider that the observed average growth of real wages paid to native-born college-educated labor over the period equaled about 13 percentage points,[35] then our accounting exercise attributes one-fourth of wage growth to H-1B-driven increases in S&E. As S&E are the drivers of technological growth, which is likely responsible for most of the long-run wage

[35] This combines the decade of the 1990s, in which there was rapid economic growth, with the first decade of the 2000s, in which the real wage slightly declined.

growth of college-educated workers, it is plausible that foreigners, who drove almost half of the S&E increase, are responsible for one-fourth of wage growth.

Next, suppose that the annual number of H-1B entrants were to rise by 50,000, as has been proposed by the current comprehensive immigration reform (from 65,000 to 115,000). That rise represents about 0.03 percent of the labor force and could, therefore, add an extra 2 percentage points to income per capita of native college-educated workers over the next 20 years (or roughly increase yearly wages by $1,330 per college-educated worker over that period). Those productivity effects, driven by the growth of S&E and their contribution to technology innovation, are certainly nonnegligible and persistent contributors to economic prosperity of the native economy.

8. CONCLUSIONS AND DISCUSSION

Many countries compete for internationally mobile S&E, believing that such labor will generate economic benefits for the domestic populace. This study provides a serious attempt at quantifying the effect of foreign S&E on the labor outcomes of native workers in local receiving economies. We analyze the United States and Canada, two countries that received a large majority of all international emigrant S&E. We then provide an evaluation of the H-1B program since its inception to the current decade and estimate the effect for the U.S. economy if the government approved an increase of 50,000 H-1B visas per year over the next 20 years.

Our analysis relies on an identification strategy exploiting historical variation in immigrant science and engineering populations across cities, combined with national-level changes in skilled-labor flows. For the United States, we use cross–metropolitan area variation and the aggregate H-1B visa flows as an exogenous shock, whereas in Canada we use cross-province variation and aggregate inflows of S&E driven by the FSWP. We test our instruments against a battery of robustness checks to see that they are robust to several controls and to validate that they are supply-driven determinants of S&E flows across cities. We find that foreign S&E increase the wages paid to college-educated natives without creating detrimental labor-market effects for any other group of native-born workers (non-college-educated overall, or separately considering high school graduates and high school dropouts). Our analysis of Canada shows

possible signs that the inflow of foreign S&E may have contributed to labor-market polarization.

Political rhetoric and debate over immigration reform in the United States have been building momentum and were further elevated in 2013 by the passage of Senate Bill 744, a comprehensive proposal to reform immigration policy. Our methodology allows us to perform policy evaluation for skilled-immigration programs, namely, the H-1B program in the United States. We find that H-1B visas awarded from the program's inception in 1990 through 2010 resulted in an augmented science and engineering labor pool that raised the growth rate of wages paid to native-born college-educated workers by 3.5 percentage points—a figure accounting for roughly a fourth of their total wage growth over that period. If policymakers approve an additional 50,000 H-1B visas per year, as has been suggested by recent proposals, they should generate an extra 2 percentage points in wage growth for native-born college-educated workers over the next 20 years. We also find that, with regard to wage effects, those benefits far exceed those generated by other commonly advocated policies for promoting growth, such as attracting high-growth industries into a city or attracting college students. We also found some evidence that those beneficial wage effects are particularly strong in metro areas with fast-growing sectors.

Although we made no attempt to quantify the potential costs of the considered growth-promoting policies—an important part of full policy evaluation—there are three reasons that the cost of expanding the H-1B visa program may actually be negligible, could generate additional benefits (besides growth), or both. First, the cost of administering and expanding the H-1B visa program, linked mainly to labor verification and processing of forms, is essentially fully covered by the fees paid by companies that sponsor the workers. It does not entail further cost for the government. To the contrary, efforts to attract sectors or companies are usually achieved with tax cuts that cost the state revenue, while establishing or expanding new universities may also have a large investment cost. Second, temporary immigrants who work in science and engineering receive high incomes and pay more in taxes than they receive in public benefits, hence representing a net return for the federal and state governments. Third, an increase in the overall H-1B program could possibly benefit all local economies in the United States and could also have some larger national spillover effects.

Other city-level growth-promoting strategies, instead, might be a zero-sum game. For example, city policies to attract fast-growing industries often divert economic activity from other localities. Efforts to increase local university presence can easily lead to graduates who move and contribute economically elsewhere.

More research and quantitative assessment in this area are certainly needed to continue to provide insight into the importance of immigration policy's economic effects. In particular, understanding the effects of international labor flows would be improved with further examination of other countries that have similarly received high inflows but have institutional structures that differ from the United States.

REFERENCES

Altonji, Joseph G., and David Card. 1991. "The Effects of Immigration on the Labor Market Outcomes of Less-Skilled Natives." In *Immigration, Trade, and the Labor Market*, ed. John M. Abowd and Richard B. Freeman, pp. 201–34. Chicago: University of Chicago Press.

Auriol, Laudeline. 2007. "Labour Market Characteristics and International Mobility of Doctorate Holders: Results for Seven Countries." OECD STI Working Paper 2007/2, Organization for Economic Cooperation and Development, Paris.

_____. 2010. "Careers of Doctorate Holders: Employment and Mobility Patterns." OECD STI Working Paper 2010/4, Organization for Economic Cooperation and Development, Paris.

Autor, David H. 2010. "The Polarization of Job Opportunities in the U.S. Labor Market: Implications for Employment and Earnings." Center for American Progress and the Hamilton Project, Washington, DC.

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. 2006. "The Polarization of the U.S. Labor Market." *American Economic Review* 96 (2): 189–94.

Bartik, Timothy J. 1991. "The Effects of Metropolitan Job Growth on the Size Distribution of Family Income." Upjohn Working Papers and Journal Articles 91-06, W. E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Borjas, George J., Jeffrey Grogger, and Gordon H. Hanson. 2012. "Comment: On Estimating Elasticities of Substitution." *Journal of the European Economic Association* 10 (1): 198–210.

Card, David. 2001. "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration." *Journal of Labor Economics* 19 (1): 22–64.

Citizenship and Immigration Canada. 2010. "Evaluation of the Federal Skilled Worker Program."

Docquier, Frederic, Calgary Oden, and Giovanni Peri. Forthcoming. "The Labor Market Effects of Immigration and Emigration in OECD Countries." *Economic Journal*.

Ellison, Glenn, and Edward L. Glaeser. 1999. "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?" *American Economic Review* 89 (2): 311–16.

Franzoni, Chiara, Giuseppe Scellato, and Paula Stephan. 2012. "Foreign-Born Scientists: Mobility Patterns for Sixteen Countries." NBER Working Paper no. 18067, National Bureau of Economic Research, Cambridge, MA.

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti. 2010. "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings." *Journal of Political Economy* 118 (3): 536–98.

Griliches, Zvi. 1992. "The Search for R&D Spillovers." *Scandinavian Journal of Economics* 94 (Supplement): S29–S47.

Hunt, Jennifer. 2011. "Which Immigrants Are Most Innovative and Entrepreneurial? Distinctions by Entry Visa." *Journal of Labor Economics* 29 (3): 417–57.

Hunt, Jennifer, and Marjolaine Gauthier-Loiselle. 2010. "How Much Does Immigration Boost Innovation?" *American Economic Journal: Macroeconomics* 2 (2): 31–56.

Hunter, Rosalind S., Andrew J. Oswald, and Bruce Charlton. 2009. "The Elite Brain Drain." *Economic Journal* 119: 231–51.

Iranzo, Susana, and Giovanni Peri. 2009. "Schooling Externalities, Technology, and Productivity: Theory and Evidence from U.S. States." *Review of Economics and Statistics* 91 (2): 420–31.

Jones, Charles I. 1995. "Time Series Tests of Endogenous Growth Models." *Quarterly Journal of Economics* 110 (2): 495–525.

Kerr, William R., and William F. Lincoln. 2010. "The Supply Side of Innovation: H-1B Visa Reforms and U.S. Ethnic Invention." *Journal of Labor Economics* 28 (3): 473–508.

Lofstrom, Magnus, and Joseph Hayes. 2011. "H-1Bs: How Do They Stack Up to U.S. Born Workers?" Discussion Paper no. 6259, Institute for the Study of Labor, Bonn, Germany.

Moretti, Enrico. 2004a. "Estimating the Social Return to Higher Education: Evidence from Longitudinal and Repeated Cross-Sectional Data." *Journal of Econometrics* 121 (1): 175–212.

_____. 2004b. "Workers' Education, Spillovers and Productivity: Evidence from Plant-Level Production Functions." *American Economic Review* 94 (3): 656–90.

_____. 2010. "Local Multipliers." *American Economic Review* 100 (2): 373–77.

_____. 2011. *The New Geography of Jobs*. New York: Houghton Mifflin Harcourt.

National Science Board. 2012. *Science and Engineering Indicators: 2012*. Arlington, VA: National Science Foundation.

OECD (Organization for Economic Cooperation and Development). 2010–2012. *International Migration Outlook*. Paris: OECD.

Ottaviano, Gianmarco I. P., and Giovanni Peri. 2006. "The Economic Value of Cultural Diversity: Evidence from U.S. Cities." *Journal of Economic Geography* 6 (1): 9–44.

_____. 2012. "Rethinking the Effect of Immigration on Wages." *Journal of the European Economic Association* 10 (1): 152–97.

Peri, Giovanni. 2011. "Rethinking the Area Approach: Immigrants and the Labor Market in California." *Journal of International Economics* 84 (1): 1–14.

Peri, Giovanni, and Chad Sparber. 2011. "Highly Educated Immigrants and Native Occupational Choice." *Industrial Relations* 50 (3): 385–411.

Scellato, Giuseppe, Chiara Franzoni, and Paula Stephan. 2012. "Mobile Scientists and International Networks." NBER Working Paper no. 18613, National Bureau of Economic Research, Cambridge, MA.

U.S. Citizenship and Immigration Services. 1998–2011. "Characteristics of H-1B Specialty Occupation Workers." Department of Homeland Security, Washington, DC.

U.S. Department of Homeland Security. 2011. *Report on Specialty Occupation (H-1B) Workers*. Washington, DC: DHS.

U.S. Department of Labor. 2011. "Office of Foreign Labor Certification Annual Report: October 1, 2010–September 30, 2011." Washington, DC.

U.S. Department of State. 2011. "Nonimmigrant Visa Statistics Classes of Nonimmigrants Issued Visas—FY 1987–2010." Washington, DC, Detail Table.

APPENDIX. CANADIAN DATA AND FIRST-STAGE ESTIMATES

The source of our Canadian data on occupations, employment, wages, rents, age, education, and nativity of individuals are provided by Statistics Canada. We use the Canadian Census Public Use Microdata File on Individuals (PUMFI)[36] in 1981, 1991, 1996, 2001, and 2006. The 1981 PUMFI is a 2 percent sample, and the 1991–2006 files are roughly 3 percent samples.[37] To protect the confidentiality of surveyed individuals, the Census Bureau recoded PUMFI variables to highly aggregate resolutions. In particular, PUMFIs identify only an individual's province[38] of residence and sometimes the census metropolitan area (CMA) in regions with large populations, thereby greatly reducing the available geographic variation. Thus, the geographical level at which we can aggregate our data is constrained by the available geographic codes in the PUMFIs.

We use data on 17 geographic areas that can be consistently identified over the period 1981–2006. Fourteen of those are metropolitan areas, including some large CMAs (Toronto, Montreal, Vancouver, Calgary, and Ottawa), and smaller CMAs (Quebec City, Hull, St. Catherines–Niagara, London, Hamilton, Winnipeg, Edmonton, Halifax, and Kitchener). We also include provinces that have no identified CMAs and hence are considered as such (Saskatchewan, New Brunswick, and Newfoundland and Labrador). That concatenation of provinces and CMAs into geographic areas is due largely to data constraints, but it does afford an interpretation of those geographic areas as labor markets. Metropolitan areas are compact economic units, and the three small provinces that we add can also be considered as relatively self-contained. Our study refers to those 17 geographic areas or local labor markets.

[36] These micro data of the Canadian Census are the only ones available to non-Canadians.

[37] The file for 1991 is a 3 percent sample; the 1996 file is a 2.8 percent sample; and the 2001 and 2006 files are 2.7 percent samples.

[38] Canada officially has 10 provinces: Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec, and Saskatchewan. The PUMFI identifies them and defines an 11th province, which is the aggregation of the three territories: Northwest Territories, Nunavut, and Yukon.

To construct the S&E occupations, we first rank occupations on their intensity of employment with a postsecondary degree in a science and engineering field.[39] Then, we classify S&E as the top-ranked in science-major intensity that constitute around 10 percent of employment in each year. Thus, a scientist or engineer is an individual who works in one of those science and engineering occupations.

We construct a variable that captures the supply-driven increase of foreign S&E in each of the 17 geographic areas between 1991 and 2006, similar to what we did for the United States, as described in Section 4.2. A key difference for Canada, however, is that we construct the aggregate growth of immigrant S&E by nationality in each period using Canadian census data (on all immigrants) rather than the H-1B data.

A.1 First Stage for Canadian Data

The growth of foreign S&E in a Canadian geographic area was driven in part by immigrant preferences, affected by the historical distribution of immigrants in 1981, but also by demand and productivity increases. In this section, we analyze how the supply-driven increase in foreign S&E affected the net observed increase in foreign S&E across Canadian geographic areas. We estimate specification (5) in the text using the imputed foreign S&E as the explanatory variable.

The coefficient of interest is b_1, which measures the effect of imputed–foreign S&E inflows on the actual increase in foreign S&E workers. It is reported in Table A3 for several specifications. We include t = 1991, 1996, 2001, 2006 so that the changes Δ refers to the periods 1991–1996, 1996–2001, and 2001–2006.

In specification (1) of Table A3, we do not include any fixed effects. In specification (2), we include period fixed effects (this is the basic specification), and in specification (3) we include the more demanding province and period fixed effects. If we consider that we have only 51 area-by-period observations, then the last specification with 13 fixed effects is very demanding. The effect of supply-driven S&E is always significant at the 5 percent level, and its value is always

[39] Postsecondary majors are identifiable from 1991 to 2006, and science and engineering fields are defined as agricultural science and technologies, engineering, applied sciences and related technologies and trades, nursing and nursing assistance, other health professions, and mathematics and physics.

close to 1. That finding implies that an increase in supply-driven immigrant S&E by 1 percentage point of employment increases the actual foreign S&E share of employment by 1 percentage point. Note that the *F*-statistic of 17.93 in the basic specification is well above the critical value for weak instrument tests (usually set around 10). Only when we include province effects does the imputed variable (though still significant) become less powerful in predicting foreign S&E (*F*-statistic equal to 4.91).

In column (4) of Table 4, we explore whether the imputed foreign S&E variable had a significant effect on the total increase in all S&E (including native workers). Although less powerful than in predicting foreign S&E, the imputed variable has a significant effect (at the 5 percent level) on the growth of total S&E (as percentage of the employment).

Table A1
S&E Classifications

A. O*NET Occupations Classified as Science and Engineering Jobs	
Actuaries	Licensed practical nurses
Aerospace engineers	Machine operators
Agricultural and food scientists	Machinery maintenance occupations
Airplane pilots and navigators	Machinists
Atmospheric and space scientists	Management analysts
Automobile mechanics	Managers of medicine and health occupations
Biological scientists	Mathematicians and mathematical scientists
Biological technicians	Mechanical engineers
Boilermakers	Medical scientists
Carpenters	Millwrights
Cementing and gluing machine operators	Not-elsewhere-classified engineers
Chemical engineers	Operations and systems researchers and analysts
Chemical technicians	Other science technicians
Chemists	Petroleum, mining, and geological engineers
Chief executives and public administrators	Physicists and astronomers
Civil engineers	Plant and systems operators, stationary engineers
Computer software developers	Plasterers
Construction inspectors	Plumbers, pipe fitters, and steamfitters
Drafters	Power plant operators
Drillers of oil wells	Programmers of numerically controlled machine tools
Drilling and boring machine operators	Rollers, roll hands, and finishers of metal
Electrical engineer	Sales engineers
Elevator installers and repairers	Secondary school teachers
Engineering technicians	Statistical clerks
Explosives workers	Supervisors of agricultural occupations
Farm managers, except for horticultural farms	Supervisors of mechanics and repairers

(continued)

Table A1
(continued)

A. O*NET Occupations Classified as Science and Engineering Jobs	
Geologists	Surveyors, cartographers, mapping scientists
Heating, air conditioning, and refrigeration mechanics	Tool and die makers and die setters
Industrial engineers	Veterinarians
Lathe, milling, and turning machine operatives	Water and sewage treatment plant operators

B. Occupations Classified as Requiring a Science and Engineering College Major	
Actuaries	Mechanical engineers
Aerospace engineers	Medical scientists
Agricultural and food scientists	Metallurgical and materials engineers
Airplane pilots and navigators	Not-elsewhere-classified engineers
Atmospheric and space scientists	Occupational therapists
Biological scientists	Optometrists
Biological technicians	Other health and therapy
Chemical engineers	Petroleum, mining, and geological engineers
Chemical technicians	Pharmacists
Chemists	Physical therapists
Civil engineers	Physicians
Clinical laboratory technologies and technicians	Physicians' assistants
Computer software developers	Physicists and astronomers
Computer systems analysts and computer scientists	Podiatrists
Dentists	Psychologists
Dietitians and nutritionists	Sales engineers
Electrical engineers	Speech therapists
Geologists	Subject instructors (HS/college)
Industrial engineers	Therapists
Management analysts	Veterinarians
Mathematicians and mathematical scientists	Vocational and educational counselors

(continued)

Table A1
(continued)

C. College Majors Classified as Science and Engineering Fields	
Aerospace Engineering	Industrial and Organizational Psychology
Animal Sciences	Industrial Production Technologies
Applied Mathematics	Information Sciences
Architectural Engineering	Library Science
Astronomy and Astrophysics	Materials Engineering and Materials Sciences
Atmospheric Sciences and Meteorology	Materials Science
Biochemical Sciences	Mathematics
Biological Engineering	Mathematics and Computer Science
Biology	Mechanical Engineering
Biomedical Engineering	Mechanical Engineering Related Technology
Botany	Medical Technologies Technicians
Chemical Engineering	Metallurgical Engineering
Chemistry	Microbiology
Civil Engineering	Military Technologies
Clinical Psychology	Mining and Mineral Engineering
Cognitive Science and Biopsychology	Miscellaneous Biology
Communication Disorders Sciences	Miscellaneous Engineering
Computer and Information Systems	Miscellaneous Engineering Technologies
Computer Engineering	Miscellaneous Psychology
Computer Information Management	Molecular Biology
Computer Networking and Telecommunication	Multi-disciplinary or General Science
Computer Programming and Data Processing	Naval Architecture and Marine Engineering
Computer Science	Neuroscience
Counseling Psychology	Nuclear Engineering
Ecology	Nuclear, Industrial Radiology, and Bio
Educational Psychology	Nutrition Sciences
Electrical Engineering	Oceanography

(continued)

Table A1
(continued)

C. College Majors Classified as Science and Engineering Fields	
Electrical Engineering Technology	Petroleum Engineering
Engineering and Industrial Management	Pharmacology
Engineering Mechanics, Physics	Pharmacy, Pharmaceutical Sciences, and Treatment Therapy Professions
Engineering Technologies	Physical Sciences
Environmental Engineering	Physics
Environmental Science	Physiology
Food Science	Plant Science and Agronomy
General Engineering	Psychology
General Medical and Health Services	Social Psychology
Genetics	Soil Science
Geological and Geophysical Engineering	Statistics and Decision Science
Geology and Earth Science	Transportation Sciences and Technologies
Geosciences	Treatment Therapy Professions
Health and Medical Preparatory Programs	Zoology
Industrial and Manufacturing Engineering	

Table A2

H-1B Visas Composition by Place of Origin

Region	**Percentage of Total, 1990–2000**	**Percentage of Total, 2000–2010**
Africa	3	2
Canada	0	0
China	5	7
Eastern Europe	5	4
India	45	47
Japan	3	3
Korea	1	3
Mexico	3	4
Oceania	2	1
Philippines	3	3
Rest of Americas	5	8
Rest of Asia	1	9
Western Europe	16	11
Other	0	0
Total H-1B visas	709,505	1,321,028

Table A3

Predictive Power of the Imputed Increase in Foreign S&E, Canadian Data (17 Canadian geographic areas; 1991–1996, 1996–2001, 2001–2006)

	(1)	(2)	(3)	(4)
Dependent Variable:	**Change in Foreign S&E as % of Initial Employment**	**Change in Foreign S&E as % of Initial Employment**	**Change in Foreign S&E as % of Initial Employment**	**Change in Total S&E as % of Initial Employment**
Imputed change in foreign S&E	1.00***	1.16***	1.28**	1.32**
	(0.24)	(0.31)	(0.58)	(0.52)
Observations	51	51	51	51
Period Effects	No	Yes	Yes	Yes
Province Effects	No	No	Yes	No
F-test of the Coefficient	17.93	14.05	4.91	6.43

Note: Each column reports coefficients from a separate regression. The units of observations are 17 Canadian metro areas and provinces over the periods 1991–1996, 1996–2001, and 2001–2006. The dependent variable is described at the top of the column. Each regression includes the 1980 native S&E and share of employment as a control. The explanatory variable is always the imputed growth of foreign S&E workers as a percentage of initial employment. ***, ** and * indicate significance at 1, 5, and 10 percent level, respectively.

Comment

Daniel Shoag

Few papers attempt to address real-world policy questions and to demonstrate important theoretical concepts at the same time. Giovanni Peri, Kevin Shih, and Chad Sparber have done that in their intriguing paper on foreign scientists, H-1B visas, and economic growth. They find that those foreign science and technology workers have very large positive local externalities, with each one raising the incomes of his or her neighbors somewhere in the tens—or perhaps even hundreds—of thousands of dollars per year. Such large externalities obviously have important implications for H-1B visa limits and immigration policy. The authors also provide evidence of substantial local knowledge spillovers, a concept that is frequently discussed in the urban economics literature but that has mostly proved empirically intractable.

Although I think highly of this paper, as a discussant I am obliged to express some reservations. As with most empirical papers, the identification assumptions are imperfect, and omitted variable bias might inflate the externalities being estimated. Extrapolating from the cross-sectional results to aggregate effects, as the authors do in their discussion, requires a leap (or several) beyond the data. I discuss both of those issues in depth below. Still, I think theirs is an important contribution; I am very glad to have read it and I encourage you to read it too.

Identification Assumptions

Immigration patterns are related to city-level growth rates in complex, forward-looking ways. The major contribution of this paper is in wrestling with that problem seriously by taking an instrumental

Daniel Shoag is an assistant professor of public policy in the Kennedy School of Government at Harvard University.

variables approach. The paper constructs a predicted inflow instrument for each city using the city's initial share of immigrants from each country and the aggregate flows from those countries to the United States. That instrument is a good predictor of actual city-level inflows because immigrants from specific countries tend to move to the same locations over time. In other words, Chinese immigrants tend to move to cities with large Chinese communities, Moroccan immigrants tend to move to cities with large Moroccan communities, and so on. The total number of Chinese immigrants to the United States is a good predictor of immigration to cities with large Chinese communities.

For the instrument to be 100 percent valid, though, it must be orthogonal to other changes in the local economy. If aggregate country-level flows were determined by national H-1B visa policy, that would be a natural assumption. Unfortunately, for most of the data examined in this paper, that is not the case. Until 2004, aggregate country-level flows seem to be driven more by the complex, forward-looking decisions of individual migrants than by arbitrary regulations.

The predictive power of the instrument stems from the fact that immigrants from one country are disproportionately likely to move to a subset of cities. It makes sense, then, that those aggregate flows would also be disproportionately influenced by the economic prospects of those cities. If that's the case, the instrument no longer reliably captures only the causal effect of immigration. For example, if Moroccan immigrants are concentrated in five cities, aggregate migration from Morocco will depend on the average economic shocks occurring in those cities. Using the aggregate flow as an instrument might still produce biased results, even if the aggregate flow is largely independent of the outcome of any single city.

Does that confounding effect explain all of the results the paper documents? I highly doubt it. Still, I think it's important to note that there are potential sources of bias working in the same direction as the estimated effect.

Setting aside those concerns, the analysis in the paper also relies on the assumption that initial settlement patterns were not correlated with subsequent "technological and demand shocks." If initial patterns are correlated with those shocks, the analysis, again, can't fully isolate the causal effect of migration.

The paper notes that the percentage of foreign-born scientists and engineers is not randomly distributed across locations. That

percentage is, intuitively, highly correlated with the total percentage of foreign-born residents. In my clumsy fumbling with analogous Integrated Public Use Microdata Series data, I find both the share of foreign scientists and the share of foreign-born workers in 1980 were highly correlated with the total share of scientists,[1] college attainment, and income.

We know that workers with higher incomes and better education received a different set of shocks over the past 30 years than their lower-income, less educated contemporaries. Skill-biased technological change, import competition from China, and tax code changes all affected skilled and unskilled cities differently. Given the nonrandom assignment of foreign scientists, those effects could also generate omitted variable bias absent sufficient controls.

The authors try valiantly to address that problem, but I am concerned that the specifications used in this paper cannot completely control for the differential shocks. The type of bias I raised above calls for city-level fixed effects, evenly spaced time intervals and controls, and spatial correlation corrections.[2] The most convincing possible specifications would go further still and control for different time effects for skilled and unskilled cities and other interaction terms. Placebo tests could also demonstrate that their instrument does not correlate with industry or demographic predicted wage and employment growth. I doubt the data are sensitive enough to allow for those kinds of controls and tests in every regression, and the paper does a good job within the bounds of that constraint. Ultimately, though, the goal is to compare wage growth in otherwise identical places, one with foreign scientists and one without. I am worried that the results are sometimes driven by comparing changes in places that systematically differ along other dimensions.

No empirical study is ever perfect, and my concerns about identification in this exercise could be leveled against many papers. Still, I think some caution should be exercised in attributing the full effect of the wage growth to the immigration mechanism alone.

[1] Surprisingly, the paper finds no correlation between the percentage of foreign scientists and native scientists under one classification of scientific workers. Under the other definition, and in my replication attempts, there is an intuitive positive association.

[2] The authors do include some of these controls in robustness tests. They reassuringly have little effect on the results.

Interpretation

The paper shows that cities receiving more foreign science and technology workers, for whatever reason, experience faster wage growth than other cities. That is a relative, not absolute, effect. The data cannot say whether the effect occurs because cities with more immigrants grew faster or whether the "other" cities grew more slowly. The authors favor the first interpretation, and, in fact, tout immigration as an alternative to other "zero-sum" growth policies. I agree that this interpretation seems plausible. Even so, it is certainly possible that the latter interpretation is true. Immigration by foreign scientists may have caused industries to move or high-skilled talent to relocate in ways that magnify the local effect without increasing the overall pie. Simultaneously, the presence of foreign scientists might also affect the incentives of natives to pursue an education, thus altering the domestic skill mix. The paper does not, and perhaps cannot, address those possibilities or quantify them. As a result, it is hard to know whether the large local effects represent gains from allowing more immigration nationally or whether immigration simply creates geographic or skill-based winners and losers.

At the same time, it seems unlikely that the positive spillovers of foreign workers are primarily local. The paper imagines those spillovers as the result of immigrants "generating innovations and developing technologies." In general, it would seem like those types of advances would diffuse rapidly across cities. I have trouble assessing what the relative local gains estimated here imply about the national benefits. Could the benefits of foreign migration be an order of magnitude larger than the estimates here? Could they be several orders of magnitude larger? It is difficult to imagine, given that the local estimates alone imply that H-1B visas can account for 25 percent of the wage growth of college workers from 1990 to 2010. That said, the value of empirical work is to update one's priors, and this paper has certainly made me rethink just how important immigration policy might be.

To sum up, I think the authors have done a great job attacking a very challenging and relevant empirical problem. The paper deals with the problem carefully, and though I think there are still some issues with the analysis, the results highlight just how critical it is to design our immigration system well. I hope that this message and the thoughtful research underlying it have an effect on the policymakers who are designing that system.

Comment

Jeffrey Smith

In their fine contribution to this volume, Peri, Shih, and Sparber examine the economic effect of foreign scientists and engineers (S&E) who enter the United States via the H1-B visa program and who enter Canada via the qualitatively similar Foreign Skilled Worker Program. Doing so presents a serious challenge to the empirical researcher because both programs operate at the national level.

One natural approach would try to exploit variation over time in the number of visas issued. Several concerns militate against adopting that empirical strategy, however. First, the H1-B program began in 1990, meaning that only a relatively short time series exists with which to estimate effects. Second, the available time series provides only limited variation, and some of that variation comes in the form of changes in the nature of those eligible for H1-B visas rather than in the simple number of such visas. Third, the extent and timing of the changes we observe likely reflect endogenous policy responses to economic shocks, and so they lack the exogeneity required for straightforward interpretation.

Given those problems with relying solely on the time-series variation, Peri, Shih, and Sparber wisely choose an alternative empirical strategy. Their approach relies on variation across cities in the number of immigrant S&E to identify the effects of such immigrants on the labor-market outcomes of native workers. That strategy has a long tradition in the broader literature that examines the effects of immigrants on natives. Because immigrant S&E choose which cities to move to, rather than being randomly allocated in some vast favor to the research community, the authors cannot simply correlate

Jeffrey Smith is professor of economics and public policy at the University of Michigan and a research associate with the National Bureau of Economic Research.

The author thanks Jennifer Hunt and Caroline Theoharides for helpful comments.

native outcomes at the city level with some function of the number of current or past immigrant S&E. Instead, they rely on an instrumental variables strategy about which I will say more later.

The authors find quite large positive effects on the wages and employment of high-skill natives in both the United States and Canada. In Canada, they find some positive effects of immigrant S&E on low-skill workers as well.

The Quality of the Music Depends on the Quality of the Instruments

Following a long tradition in the literature, the authors use historical migration patterns as instruments for current migration patterns. I like to refer to this as an example of the "fine wine" theory of instrument selection, as it presumes that particular variables, like fine wine, improve with age by becoming more exogenous—and thus more pleasing to the "instrument police."

The particular instrument used here relies on historical shares of immigrant S&E from particular countries in particular cities. Those historical shares represent a valid instrument only if they do not correlate with the unobserved component of future outcomes (i.e., with the error terms in the authors' outcome equations). The authors recognize this issue and devote much of their Section 4.3 to providing evidence that the reader should not worry overmuch about it. I applaud the authors' efforts, but it remains the case that if you take their estimates seriously and project them back in time to early cohorts of immigrant S&E, then the instrumental variable condition should fail, as cities with more immigrant S&E in the past should have differentially large employment and wage growth (for the relevant groups) during the period the authors study. To avoid that conclusion, I think the authors would have to argue either that earlier cohorts had smaller effects than the cohorts they study, or that immigrant S&E effects fade out over time, which would affect the interpretation of their own estimates.

Interpretation: Heterogeneous Coefficients

We often speak of "the" effect of a program or policy, but in fact the effects of policies vary along many different dimensions. In the context of the authors' paper, I want to highlight three dimensions on which thinking more about heterogeneity in the effects of H1-B

policy (or its analogue in Canada) would add to our understanding of their results and to their policy relevance. First, should we think of those effects as short-run effects, medium-run effects, or long-run effects? Given the 5-year and 10-year calendar time windows (and I would have used the results with windows of equal length as the primary results in the paper precisely to make the interpretation clearer on this dimension), it seems that they combine all three. The change in an outcome of interest from 1 year to 10 years later captures the short-run effects of S&E who arrive late in the window, the medium-run effects of workers who arrive in the middle of the window, and the long-run effects of workers who arrive early in the window. Given this sort of omnibus parameter, more discussion of the likely relative magnitudes of the effects at different points in time following worker arrival would aid the reader and improve the cost-benefit analysis.

Second, we care about the effect that existing H1-B workers have had on average on economic outcomes because we can compare that parameter with the costs of the H1-B program and provide an answer to the question of whether the existing program passes a cost-benefit test. However, we might expect—either because of diminishing marginal quality of H1-B visa recipients or for reasons related to the shape of the production function—that the effect of additional immigrant S&E at the margin of admission might differ from the average effect of those admitted under the current policy. In thinking about an expansion in the number of H1-B visas like that suggested by the authors, the average effect at current levels likely provides an upper bound on what additional workers would bring. The policy discussion would benefit from additional attention to that aspect of the problem.

Finally, in the context of heterogeneous coefficients, we might imagine that the particular instrumental variable employed here sweeps out a nonrandom subset of migrants and, as a result, a nonrandom subset of responses to migration. The analysis in this paper estimates the effect of immigrant S&E who choose their location in the United States or Canada based on the locations (or based on factors correlated with the locations) of earlier cohorts of immigrant S&E from the same source countries. Should we expect those immigrants to have the same effects on native outcomes as a randomly selected immigrant? One can tell different stories.

Such immigrants might have larger effects if the presence of a preexisting community of immigrant S&E from the same source country speeds assimilation or, perhaps, allows them to work more efficiently because they have colleagues with the same native language and background. Conversely, perhaps the immigrants who locate in that way represent the less adventurous and innovative immigrants in their cohorts. In this case, the instrumental variable estimates obtained by the authors might understate the effect of a randomly selected foreign science or engineering worker. Whatever the story, future research should pay more attention to that aspect of heterogeneity.

Interpretation: Size Matters

Peri, Shih, and Sparber credit H1-B visa holders with causing a quarter of the total wage growth for college-educated natives over the period 1990–2010. That is a huge effect! It also seems like an implausibly large effect. Let me give one reason as to why. According to their Table 1, immigrant S&E in metro areas average about 20 percent of the science and engineering work force over that period. If we assume that domestic S&E have the same effect on wages and employment as immigrant S&E, then S&E taken together explain 125 percent of the wage growth of college-educated natives over the period. Not only would that result leave no room for factors such as technological change, it implies that the net effect of other factors is to reduce wages. Thus, in my view, to make the magnitude of the authors' estimates credible requires, for example, an argument as to why immigrant S&E have differentially positive effects on wages.

Stepping back a bit to think more broadly about the estimates suggests further concerns. To the extent that immigrant S&E create general knowledge that leads to productive innovation, we would expect knowledge, at least in the medium to long run, to have effects everywhere, not just effects that are local to specific cities. If that process occurs within 5 or 10 years, then that aspect of what immigrant S&E do gets captured not in the coefficient estimates the authors focus on, but rather it gets differenced out as part of the implicit period effect. For that reason, the authors actually understate the total contribution of immigrant S&E to U.S. output growth, something I am surprised that the authors do not say more about.

This literature would benefit from additional efforts to parse out the overall contribution into its general and city-specific components and from efforts to determine how long it takes general contributions to fully spread.

Interpretation: Microfoundations and Alternative Frameworks

As Peri, Shih, and Sparber acknowledge, their estimated effects on employment and wages combine responses by natives on a number of dimensions. When employment of native S&E increases in cities that immigrant S&E choose to live in, for example, that increase could reflect (a) changes in the schooling response of natives (for example via college students in the relevant cohorts changing their majors), (b) occupational switches by S&E currently working in other types of jobs, (c) labor force participation choices by S&E initially out of the labor force, or (d) migration of S&E from other labor markets. Microfounding (my macroeconomist colleagues assure me that "microfound" is a proper verb) the aggregate effects by estimating the relative importance of those pathways would aid in understanding how the estimated effects come about.

Such analyses have precedent in the literature. For example, Cadena (2013) looks at migratory responses by natives in a more general immigration context; Jackson (2013) considers the response of natives' educational investments to immigrant shocks; and Cortes (2008) looks at the related issue of the effects of immigrants on price levels. All three papers operate within the city-level analytic framework employed by Peri, Shih, and Sparber.

To the extent that migration of native S&E between cities plays an important role, the framework here starts to wobble a bit, for it is ill-suited to either the interpretation or estimation of such general equilibrium effects. In that regard, further work using alternative frameworks that explicitly incorporate general equilibrium considerations, such as the search framework employed by Chassamboulli and Palivos (2014) or the spatial equilibrium models used in the urban economics literature, such as Albouy (2009), would add great value to the literature. Even in the context of the current paper, more attention to equilibrium effects in interpreting the results would enrich our understanding of the findings.

Bottom Lines

In sum, the authors deserve credit for tackling an important substantive problem with great policy relevance, despite the many difficulties with which it confronts the empirical researcher. And they have made some useful headway on the problem and have provided some valuable evidence in doing so. I think much work remains in this literature; thus, I have suggested some directions I think worth pursuing in the course of my comments. At the same time, I already agreed with the authors' policy proposal before reading the paper, but reading it has strengthened that agreement in my mind.

I would like to close with one last point, which relates to the odd welfare economics of migration. Economists have the praiseworthy habit of worrying about the well-being of people who live outside the arbitrary geographic boundary of their country of origin. Importing that concern into this literature, which focuses almost exclusively on the well-being of natives in the immigrant-receiving country, would make it intellectually deeper, dispel any odor of nationalism, and make its policy conclusions more relevant to those who think everyone matters, even foreigners.

REFERENCES

Albouy, David. 2009. "The Unequal Geographic Burden of Federal Taxation." *Journal of Political Economy* 117 (4): 635–67.

Cadena, Brian. 2013. "Native Competition and Low-Skilled Immigrant Inflows." *Journal of Human Resources* 48 (4): 910–44.

Chassamboulli, Andri, and Theodore Palivos. 2014. "A Search-Equilibrium Approach to the Effects of Immigration on Labor Market Outcomes." *International Economic Review* 55 (1): 111–29.

Cortes, Patricia. 2008. "The Effect of Low-Skilled Immigration on U.S. Prices: Evidence from CPI Data." *Journal of Political Economy* 116 (3): 381–422.

Jackson, Osborne. 2013. "Does Immigration Crowd Natives Into or Out of Higher Education?" Unpublished manuscript, Northeastern University, Boston.

Evaluating Policies to Prevent Another Crisis: An Economist's View

Paul S. Willen

ABSTRACT

I consider four policies created to address the financial crisis: (a) the ability-to-repay requirement in mortgage underwriting, (b) reform of rating agency compensation, (c) risk retention in securitization, and (d) mandatory loan renegotiation. I ask whether economic theory tells us that those policies can improve on the market. I argue that policies a, b, and c are likely to reduce welfare versus the market, and only policy d has the potential to increase it.

Paul S. Willen is a senior economist and policy adviser in the research department of the Federal Reserve Bank of Boston and a faculty research fellow at the National Bureau of Economic Research.

The author thanks Cato Papers *organizer Jeff Miron for taking a chance on a somewhat speculative project and Parag Pathak and Neng Wang for thoughtful discussions at the conference. Helpful comments and suggestions from Alberto Bisin, Ryan Bubb, Chris Foote, Kris Gerardi, and Bob Triest are gratefully acknowledged.*

The views expressed in this paper are the author's and do not necessarily reflect those of the Federal Reserve Bank of Boston, the Federal Reserve System, or the Federal Open Market Committee.

Evaluating Policies to Prevent Another Crisis: An Economist's View

1. INTRODUCTION

> Economic analysis does have one important benefit, which is that it can help kill ideas that are completely logically inconsistent or wildly at variance with the data. This insight covers at least 90 percent of proposed economic policies.
>
> —Ben Bernanke, June 2, 2013[1]

In the wake of the financial crisis and recession that started in 2008, policymakers instituted legal and institutional changes with the goal of preventing a recurrence. In this paper, I evaluate four of those policies and ask the central question of classical economics: do these government policies improve upon the market?

Let me illustrate the point of this paper using mandatory risk retention, a key provision of the 2010 Dodd-Frank Wall Street Reform and Consumer Protection Act. Mandatory risk retention specifies that firms that securitize mortgages must retain not less than 5 percent of the credit risk for any security they issue backed by mortgages. The idea is that the 5 percent retained risk exposes the original lender to losses when borrowers default, thereby inducing the lender to exert more effort in underwriting loans, which, in turn, reduces losses to investors.

Is welfare higher in an economy with government-mandated risk retention than it would be without it? The answer is unclear. If the benefit of fewer defaults to the investor exceeds the cost of retained risk to the lender, why wouldn't the lender volunteer to retain the risk? Why would government need to force investors and lenders to do something that is already in their interests? The purpose of this paper is to answer those questions. Surprisingly, in the literally thousands of

[1] Ben S. Bernanke, "The Ten Suggestions" (speech at the Baccalaureate Ceremony, Princeton University, Princeton, NJ, June 2, 2013), http://www.federalreserve.gov/newsevents/speech/bernanke20130602a.htm.

papers, reports, and op-eds written about the role of securitization in the crisis, no one has really tried to answer them. Policymakers and, surprisingly, most economists have been satisfied with the argument that because risk retention leads to more effort at underwriting loans and such effort is good, risk retention is good policy.

In what follows, in addition to risk retention (Section 4), I discuss the ability-to-repay requirement (Section 2), reform of rating agency compensation (Section 3), and mandatory loan renegotiation (Section 5), each time asking the same question: does this policy improve on the market?

As with any policy analysis in economics, the starting point is the first theorem of classical welfare economics, the "First Welfare Theorem" hereafter, which states that under standard assumptions, market equilibrium is Pareto optimal. That means that if we try to reallocate resources, the only way we can make one person better off is by making someone else worse off. The necessary conditions for the First Welfare Theorem to hold are many, but the following are the ones we care about here:

- All market participants have the same ("symmetric") information.
- All market participants are rational.
- There are no externalities.

For government policies to work, one of those assumptions must fail.

Two pieces of economic jargon help here. If the First Welfare Theorem does not hold in an economy, then we say that there is a market failure in the economy. The gap between the optimal allocation (one with resources reallocated) and market equilibrium is known as the deadweight loss of the market failure. Policy analysis in classical economics essentially consists of the search for deadweight losses. The existence of a deadweight loss is a necessary condition for effective policy, and the net reduction of deadweight loss is a sufficient condition.

In Sections 2–5, the main focus is on whether failure of the first assumption, symmetric information, is enough to justify the policies. Those sections form the core of the paper because all the policies I discuss involve some form of asymmetric information. For example, the economic issue in risk retention is that investors (or anyone else providing funds to a lender, including depositors or shareholders) cannot observe how much effort the lender puts into underwriting a loan. That asymmetric information does lead to a deadweight loss in

equilibrium, but I show that, under standard assumptions, mandatory risk retention actually makes the deadweight loss worse.

For all of the policies except loan renegotiation, I argue that asymmetric information alone cannot justify government intervention. In all three cases, a more plausible argument for government intervention is that market participants are irrational. In my view, irrationality is at the heart of many of the bad decisions that caused the crisis, but the policies described fail to address the relevant problem: unrealistic beliefs about house price appreciation.

In Section 6, I turn to externalities. Foreclosure externalities could provide justification for any of the four policies because all, in theory, reduce the number of foreclosures. I draw attention to the distinction between physical externalities, which always generate deadweight losses, and pecuniary externalities, which, under standard assumptions, do not.

There exists, for all practical purposes, a theorem in economics that—for any proposed economic policy—there is some cocktail of market imperfections that can justify the policy. Geanakoplos and Polemarchakis (1986), for example, show that with multiple goods and incomplete markets, a government policy always exists that can increase welfare. But precisely because it is so easy to come up with such examples, the null hypothesis in economics has been that government policy cannot improve on the market. To show that a policy will work, one must have robust evidence of a market imperfection and a clear logic for how the proposed policy addresses it. I argue herein that only one of the proposed policies, mandatory loan modifications, plausibly passes that test.

One may read the results in this paper as saying that many of the policies developed to deal with the crisis are welfare reducing. But I think it is more accurate to say that the policies are welfare reducing in an economy composed of rational individuals in which foreclosures generate no physical externalities. As I explain, failure of either of those conditions could make all those policies welfare improving. However, to make policy work, economists cannot simply say that people are irrational or that externalities exist. If one believes, for example, that investors in subprime securities were irrational, then we need to know how they were irrational and how important that irrationality was. So, in a sense, this paper is not a criticism of policy but rather a call for better models and better research.

2. ABILITY TO REPAY

Title XIV, Subtitle B, Section 1411, of the Dodd-Frank Act states the following:

> In accordance with regulations prescribed by the Board, no creditor may make a residential mortgage loan unless the creditor makes a reasonable and good faith determination based on verified and documented information that, at the time the loan is consummated, the consumer has a reasonable ability to repay the loan, according to its terms, and all applicable taxes, insurance (including mortgage guarantee insurance), and assessments.

On the face of it, Section 1411 seems like common sense. Would a rational lender make a loan that a borrower doesn't have a "reasonable ability to repay"? And would a rational lender make a determination without "verified and documented information"? Yet I contend that the answer to both questions is yes and that, as a result, the ability-to-repay requirement generates deadweight losses.

Let us start with the idea that a lender should make loans only when the borrower has "a reasonable ability to repay the loan." On the face of it, that requirement sounds reasonable. Why would a lender make a loan if the probability of default is high? Why would a rational borrower want such a loan? The classic subprime story is of a person who has a history of credit problems; a lot of high-interest, unsecured debt; and an equity stake in his or her home. By taking out a mortgage secured by the house, the borrower can get a much lower interest rate and in doing so improve his or her finances. To be sure, the likelihood is high that the borrower will default, but the lender compensates for that by charging a much higher interest rate than if the borrower had a clean credit history.

In the example, the borrower gets a lower interest rate than would be available without home equity collateral, and the lender gains because the benefits of the high interest rate outweigh the costs of higher default risk, so both parties gain from the transaction. In this example, by preventing lenders from making loans because the borrower lacks a "reasonable ability to repay the loan," Section 1411 creates a deadweight loss: the benefits to the borrower of getting the loan exceed the costs to the lender of producing the loan.

Now we turn to the question of verification. Even if lenders want to make risky loans, doesn't it always make sense to verify what the borrower reports? The answer, again, is not necessarily. Suppose we have a lender who is confronted with 10 observably identical borrowers who have applied for $10 loans. Four of the borrowers will default on their loans, and the lender will recover nothing. To understand the lender's problem, we turn to a great insight attributed to the pioneering 19th-century retailer John Wanamaker, who famously quipped, "Half the money I spend on advertising is wasted; the trouble is, I don't know which half." The lender here might paraphrase Wanamaker, saying, "I have 10 borrowers and 4 will default; the trouble is, I don't know which 4."

Suppose, however, that by verifying the information in the borrower's application, the underwriter can identify two of the defaulters, but suppose that the verification process costs $2.50 per loan. Should a rational lender do it? One might think the answer would be an emphatic yes. By spending $2.50, the lender can avoid a much bigger loss of $10. But in fact, the answer is no. The problem is that the lender has to pay the verification costs on all 10 loans because, of course, the point here is that he can't tell which borrowers will default, so the cost of verification is actually $25. Since verification prevents only two defaults, it saves the lender only $20, so the costs of verification outweigh the benefits, and the lender opts against verification.

2.1 A Simple Model of Underwriting

To expand on this example, consider a simple model of underwriting due to Bubb and Kaufman (2009). Suppose we have a set of borrowers. For each borrower, the lender observes a verifiable, public piece of information, x. Think of x as, for example, an index of the number of times the borrower has been delinquent on his or her current mortgage in the past year. For simplicity, assume that x is some number between 0 and 1 and that x equals the probability of default. In other words, if 100 borrowers apply for loans with $x = 0.2$, then the lender knows that 20 of the borrowers will default—but, of course, the problem is that the lender does not know which 20. We can formalize the idea of verification by supposing that if the lender pays an amount c, it can identify some fraction s of the borrowers who will default. If s is 50 percent and x is 20 percent, then the lender will learn the identity of 10 of the 20 defaulters. After verification,

the lender extends loans to the remaining 90 borrowers, knowing that 10 will default, but, again, the lender does not know which 10.

Suppose that the lender earns an interest rate of $R > 1$ if the borrower repays and faces a cost of funds of 1. The lender faces a choice: to verify or not to verify. If the lender chooses not to verify, the payoff is

$$\underbrace{(1-x)(\bar{R}-1)}_{\text{Repayers}} + \underbrace{x \times -1}_{\text{Defaulters}} = \bar{R} - 1 - \bar{R}x \tag{1}$$

If the lender chooses to verify, then the payoff is

$$\underbrace{(1-x)(\bar{R}-1)}_{\text{Repayers}} + \underbrace{sx \times 0}_{\substack{\text{Identified}\\ \text{Defaulters}}} + \underbrace{(1-s)x \times -1}_{\substack{\text{Unidentified}\\ \text{Defaulters}}} - c = \bar{R} - 1 - c - (\bar{R} - s)x \tag{2}$$

Subtracting equation (1) from equation (2) yields the condition for optimal verification:

$$s \cdot x > c \leftrightarrow x > c/s \equiv x^A$$

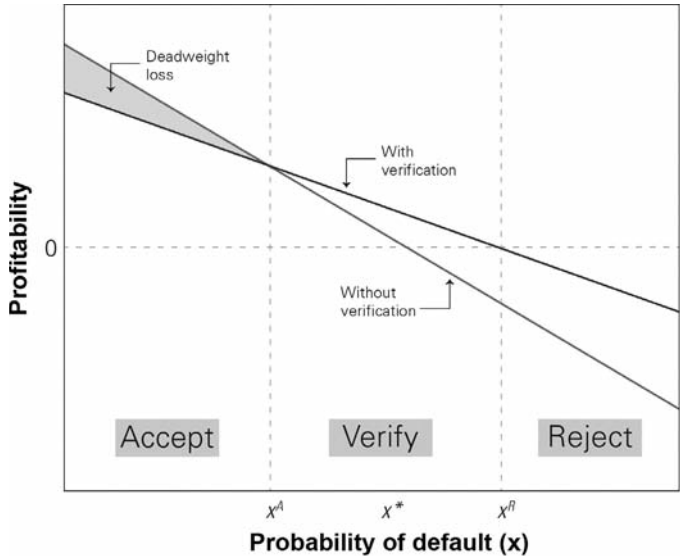
Verification makes sense only if $s \cdot x$, the losses associated with the defaulters identified by the verification, exceeds the costs, implying that if x is sufficiently low ($< x^A$), verification does not make economic sense.

To illustrate the effects of mandatory verification, I plot the payoff to verifying loans and not verifying loans as a function of x, the *ex ante* probability of default. Panel A of Figure 1 shows that we can divide borrowers into three regions. For the borrowers with $x > x^R$, lending is never profitable, even with verification. At the other extreme, the costs of screening exceed the benefits for borrowers with $x < x^A$, so the lender accepts the borrowers without screening. In the middle, the benefits of screening exceed the costs, and the lender screens. Screening clearly generates welfare benefits: for borrowers between x^A and x^R, profits are higher for the lender and, in the absence of screening, borrowers between x^R and x^* would not get loans at all. Suppose Congress now imposes mandatory verification. What happens to welfare? The shaded area on the upper left shows the resulting deadweight loss. No additional borrowers receive credit, and for the lender, the benefits of the additional screening do not justify the costs.

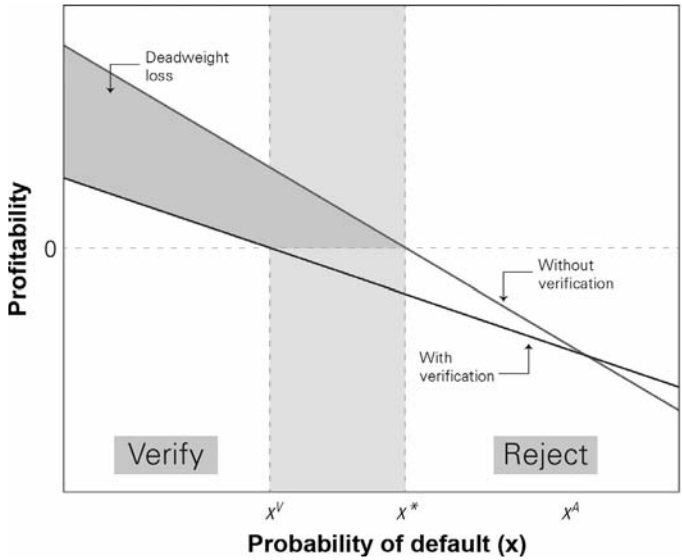
Panel A shows the deadweight loss of requiring verification when the costs of verification exceed the benefits. Panel B shows that sufficiently high verification costs prevent some borrowers from getting credit. See Section 2.1 for details.

Figure 1
A Simple Model of Loan Verification

Panel A: Deadweight loss of verification



Panel B: High verification costs



Panel B of Figure 1 shows the effect of mandatory screening when the cost of verification is even higher. Suppose, for example, that we are considering a sample of self-employed borrowers for whom measuring income is extremely complex. In this example, x^* falls below x^A, so verification dominates only for loans that are unprofitable anyway. The resulting deadweight loss is, of course, much larger, but there is an even worse aspect. Now mandatory verification means that loans between x^V and x^* are no longer profitable for the lender, and borrowers in that region no longer receive credit.

2.2 The Role of Irrationality

For lenders at least, forcing verification reduces welfare. Some critics of the lending industry have accused lenders of "not bothering" to verify income, but—as the model shows—a perfectly rational lender may choose not to verify. Proponents of the ability-to-repay rule will respond that the purpose is not to protect lenders, but to protect borrowers. But to an economist, that argument should strike as odd. Notice that in our simple model of underwriting, it is the borrower who has private information about whether he or she can repay the loan, not the lender. The purpose of the underwriting process is to maximize profits for the lender, not to maximize utility for the borrower.

Ultimately, any justification for the ability-to-repay standard has to rely on a behavioral model and argue that limiting choice increases borrower utility. Laibson (1997) has stressed the idea that when we consider alternatives to classical assumptions on preferences, limiting choice can make an individual better off. Gul and Pesendorfer (2001) consider preferences defined over sets of consumption bundles in which it is possible for a consumer to prefer a subset to the set itself. In either setup, it is possible that a borrower's utility will be higher if Congress makes it impossible for the borrower to get a loan.

My view is that the deeper problem here is that the true irrationality in the crisis involved expectations of house price appreciation. A mortgage is collateralized debt, and the assumption underlying the contract is that the value of the collateral guarantees the ability to repay because the borrower can sell the house. In 2005, both borrowers and lenders were exceptionally confident that houses would be significantly more valuable over time and very unlikely

to be less valuable. Under those circumstances, the ability-to-repay standard would have had only a minimally deterrent effect. The only way in which the ability to repay from income would be an issue is if the ability to repay by selling was not an option; because the lender believed the latter scenario to be highly unlikely, the benefits of high-interest income from subprime loans would outweigh the cost of the higher risk of a lawsuit in the event that the borrower proved unable to sell the house and lacked the income to repay the loan.

Before I conclude, it is important to stress how odd the ability-to-repay standard is. To see why, consider another important decision: hiring. If we applied a similar rule to hiring, we would require employers to carefully verify everything in a prospective job applicant's file. If the worker were hired and then subsequently fired, the worker would have the right to sue the employer if the worker found out that, for example, the employer had failed to call all the references. No such law has ever been proposed, despite the fact that the consequences of job loss are, in many cases, worse than those of default and foreclosure.

3. RATING AGENCIES

Title IX: Subtitle C of the Dodd-Frank Act specifies "Improvements to the Regulation of Credit Rating Agencies." The authors of the bill write:

> In the recent financial crisis, the ratings on structured financial products have proven to be inaccurate. This inaccuracy contributed significantly to the mismanagement of risks by financial institutions and investors, which in turn adversely impacted the health of the economy in the United States and around the world. Such inaccuracy necessitates increased accountability on the part of credit rating agencies.

As a result, they propose that in

> certain activities, particularly in advising arrangers of structured financial products on potential ratings of such products, credit rating agencies face conflicts of interest that need to be carefully monitored and that therefore should be addressed explicitly in legislation in order to give clearer authority to the Securities and Exchange Commission.

Among other things, the law directs the Securities and Exchange Commission to study the following aspects of the institutional structure of the rating agency model:

> (1) the credit rating process for structured finance products and the conflicts of interest associated with the issuer-pay and the subscriber-pay models; (2) the feasibility of establishing a system in which a public or private utility or a self-regulatory organization assigns nationally recognized statistical rating organizations to determine the credit ratings of structured finance products.

What is the basic economics here? Rating agencies analyze securities and evaluate the likelihood of credit losses and then relate that information to investors, who can, as a result, make rational decisions about whether to invest in the securities. According to the crisis consensus, the problem was that rating agencies had a conflict of interest because they were paid by the issuers of the securities they were supposed to evaluate. The result was the agencies gave optimistic ratings, which led investors to overinvest in securities and subsequently lose money.

On the face of it, that result appears to be a classic example of incentives at work, but what does economic theory say? Consider a simple model of rating agencies. Suppose there is a lemons problem in the securities market. Issuers have private information about whether a security is a peach, which is worth V_G to investors and $V_G - \delta$ to the issuer, or a lemon, which is worth $V_B < V_G$ to investors and $V_B - \delta$ to the issuer. The probability that a security is a peach is π, and we assume that the expected value of a security is $\pi V_G + (1 - \pi)V_B < (1 - \delta)V_G$. Investors know their own valuation, the valuation of the issuer, and the probability of a peach, but they do not know whether a particular security is a peach or a lemon. Akerlof (1970) showed that in such a model, the only securities traded in equilibrium are lemons. The intuition is that if the investor knows that if he pays the expected value of the security $\pi V_G + (1 - \pi)V_B$, the issuer will sell only the bad security, meaning that the investor will lose money, so the only possible equilibrium is with price V_B. Equilibrium is Pareto inefficient because the investor values the peach more than the issuer does, but trade cannot take place.

Suppose we introduce a rating agency. The issuer reveals its private information to the rating agency, which in turn can credibly announce

whether a particular security is a peach or a lemon. Now trade can occur in both securities at prices V_G and V_B, respectively. The new equilibrium is Pareto improving, as the price for the lemons stays the same, but trade now occurs for peaches, making the issuer better off because he receives V_G for something he valued at only $V_G - \delta$.

What can go wrong? The critique of rating agencies on which Dodd-Frank is based depends on the fact that the pay of the rating agency depends on the rating. Suppose we change the previous model a little and say that the rating agency is paid a share of the valuation of the security, $\gamma(V_i - V_B)$. Now the rating agency has an incentive to always report that the security is a peach because it will get paid $\gamma(V_G - V_B)$, but if it reports that the security is a lemon, it will get nothing. Now if investors pay V_G, on average they will get securities worth $\pi V_G + (1 - \pi)V_B < V_G$. According to this narrative, misaligned incentives of rating agencies can explain why investors lost money.

But that logic is flawed: the situation described is not an equilibrium. The investors in the model do not know whether a security is a peach or a lemon, but they know the incentives of the issuer and the rating agency. If the rating agency makes more money by saying that a security is a peach, then the agency will always say a security is a peach, and therefore the rating will have no meaning. As a result, the investor will treat the rating as "cheap talk," and the original Akerlof equilibrium will reemerge. Who loses? Everyone. Since all traded securities are lemons, the rating agency gets paid nothing, and both the issuer and investor lose the gains from trading peaches.

3.1 The Role of Irrationality

Of course, one could argue that investors did not understand the conflicts of interest for the agencies. But given that the firms that lost the largest sums of money on structured products were issuers of the securities, that is a hard argument to make.[2] In addition, even if one tried to argue that investors were ignorant of the corruption of the rating agencies in 2005, it is hard to see how they could remain ignorant now; yet all three rating agencies are still doing business and still rating mortgage-backed securities.

[2] Bolton, Freixas, and Shapiro (2012) construct a model with "naive" investors who do not understand the incentives of the rating agency.

3.2 Alternative Information Issues in Ratings

That said, one cannot dismiss the broader idea that bad ratings have real effects, but to do so, one needs to understand exactly what ratings were used for. Specifically, investment funds often use ratings to limit a manager's investment choices. For example, pension funds might say that the manager of the fund can invest only in AAA-rated securities.

I propose that using rating agencies to limit investment manager choice results in a different asymmetric information problem. Suppose that the returns to a pension fund depend on the unobservable effort of the manager. So following the discussion in Section 3, the manager's compensation depends on returns. However, suppose that the fund manager is risk neutral and that the investors are risk averse. By investing in riskier assets, the manager can raise the expected return and his expected pay, but at a cost of increasing risk to the investors. One solution to that problem would be to limit the fund manager to observably low-risk securities, which is exactly what funds typically do.

The implication of the fund's asymmetric information problem is that the investment manager now has an incentive to get the rating agency to give high ratings to risky products. In other words, both the issuer and the investment manager have an incentive to get high ratings for lemons. If that were the case, then the "investor pays" model would be no more likely to lead to more accurate ratings than the much-maligned "issuer pays" model.

In the end, though, the point here is that for the rating agencies to have value to issuers, they must have value to investors; otherwise investors would not pay more for a rated security than they would for an unrated security. And to have value to investors, rating agencies must be credible. In other words, the rating agencies have a strong market incentive to solve the credibility problem, and it is not clear why government needs to tell them to do it. No government agency needs to regulate *Consumer Reports* to ensure that it is objective: if it were perceived to be biased, no one would pay for the magazine.

4. RISK RETENTION

Title IX: Subtitle D of the Dodd-Frank Act, titled "Improvements to the Asset-Backed Securitization Process," specifies that firms that securitize mortgages are required to retain not less than 5 percent of

the credit risk for any security they issue that is backed by mortgages. Framers of the law included an exemption for "qualified residential mortgages," which are loans deemed to be of low risk of credit loss. The law also prohibits securitizers from hedging or transferring the credit risk that it is required to retain with respect to the assets.

Risk retention has been broadly popular across the board, earning praise from journalists, academics, and policymakers. One of the reasons so many mortgages defaulted, the argument goes, is that the lenders who made the loans were selling the loans in securities and had no reason to invest effort in underwriting the loans because they did not share in any losses when the loan defaulted. In popular parlance, lenders had no "skin in the game." If lenders knew they would lose money if loans defaulted, they would have been much more careful. With more careful underwriting, investors would not have lost money on mortgage-backed securities, and we would not have had a crisis. Keys et al. (2013), for example, write that it would be "beneficial to enforce some mandatory retention of a fraction of lower tranche by originators/underwriters to better align their interests with those of investors."

How does that view fit into our discussion of the First Welfare Theorem? Proponents of risk retention argue that securitization has an asymmetric information problem. Although one can imagine how mortgage underwriting could lead to both moral hazard and adverse selection problems, researchers and policymakers have focused on the moral hazard problem, which results from the fact that investors cannot observe how much effort the securitizer puts into screening the mortgages. As previously mentioned, the presence of asymmetric information typically leads to inefficiency, and thus in principle government policy might lead to a welfare improvement. But, as I will show now, the Dodd-Frank requirement not only fails to eliminate the deadweight loss caused by asymmetric information, but also actually inflates it.

It is important to stress here that securitization does not create the moral hazard problem but is instead a method of dealing with it. The underlying problem in financial intermediation is that savers want to lend to borrowers but need someone to help them make sure they get their money back. The moral hazard problem emanates from the fact that the savers cannot observe the effort put in by the intermediary.

Over the years, market participants have come up with many different mechanisms to deal with the moral hazard problem. Securitization is one contract—or incentive scheme as we call it below—in which the intermediary takes on very little of the risk of default. An alternative contractual mechanism is portfolio lending, in which the owners of the bank (who may not be the ones making the lending decision) take on all the risk. The problem of choosing the optimal contract consists of deciding which incentive scheme maximizes the joint surplus of the intermediary and the savers. In other words, which incentive scheme deals best with the underlying problem that the savers cannot observe the effort of the intermediary?

To understand the problem of incentives in mortgage underwriting, we consider a pool of 10 loans. If the lender puts no effort into underwriting the loans, three borrowers will default but, as always, the lender does not know which three. By investing effort, the lender can identify the problem borrowers. The column labeled "Total Effort" in Panel A of Table 1 shows that by spending $4 in effort, the lender can identify one of the problem borrowers and thus reduce defaults to two. The column labeled "Marginal Effort" shows that the cost of reducing defaults is increasing: the marginal effort required to reduce defaults from two to one costs twice as much as it does to reduce defaults from three to two. The columns labeled "Recovery" show the benefits of default prevention: for each default prevented, the total recovery of principal increases by $10.

Now suppose that we have an investor who wants to invest in mortgages, and so he offers to buy mortgages from a lender. The lender is willing to underwrite mortgages but has an outside option that will yield a profit of $69. In the language of contract theory, we call the investor "the principal" and the lender "the agent." How can the principal ensure that the agent expends the proper amount of effort? Panel B shows three possible incentive schemes that the principal can use. In each scheme, the principal promises the agent a combination of a fixed payment and an incentive payment, which is a fraction of the amount recovered. For example, Incentive Scheme 1 gives the agent a base payment of $33, and then the agent keeps 50 percent of any money recovered. For example, if the agent expends effort of $4, the agent will receive $33 + 0.5 \times 80 = \$73$, which yields a profit to the agent of $69, which is the $73 payment less the $4 expended in effort.

Table 1

A Simple Model of Moral Hazard in Mortgage Underwriting
(see Section 4 for details)

Panel A. Costs and benefits of effort

	Agent		Principal	
Defaults	**Total Effort**	**Marginal Effort**	**Total Recovery**	**Marginal Recovery**
3	0	0	70	0
2	4	4	80	10
1	12	8	90	10
0	24	12	100	10

Panel B. Choosing the optimal incentive scheme without uncertainty

		Incentive Scheme 1 Incentive: 50% Base Pay: 33				**Incentive Scheme 2 Incentive: 90% Base Pay: 0**				**Incentive Scheme 3 Incentive: 130% Base Pay: -37**			
		Agent				*Agent*				*Agent*			
Total Effort	**Defaults**	**Total Income**	**Marginal Income**	**Utility**	**Principal's Profit**	**Total Income**	**Marginal Income**	**Utility**	**Principal's Profit**	**Total Income**	**Marginal Income**	**Utility**	**Principal's Profit**
0	3	68		68	2	63		63	7	54		54	16
4	2	**73**	**5**	**69**	**7**	72	9	68	8	67	13	63	13
12	1	78	5	66	12	**81**	**9**	**69**	**9**	80	13	68	10
24	0	83	5	59	17	90	9	66	10	**93**	**13**	**69**	**7**

(continued)

Table 1
(continued)

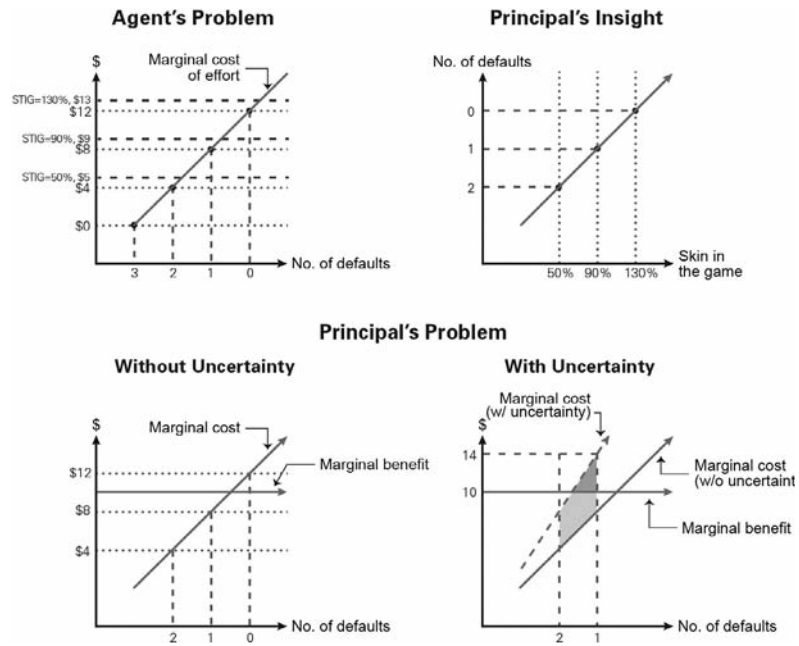
Panel C. Choosing the optimal incentive scheme with uncertainty

		Incentive Scheme 1 Incentive: 50% Base Pay: 37						**Incentive Scheme 2 Incentive: 90% Base Pay: 8**					
		Agent						*Agent*					
Total Effort	***Expected* Defaults**	**Total Income**		**Marginal Income**	**Cost of Risk**	**Utility**	**Principal's Profit**	**Total Income**		**Marginal Income**	**Cost of Risk**	**Utility**	**Principal's Profit**
		Bad	**Good**					**Bad**	**Good**				
0	3	67		4	68	68	−2	62	82		8	63	−1
4	2	**72**	**5**	**4**	**69**	**69**	**3**	70	90	9	8	68	0
12	**1**	77	5	4	66	66	8	**79**	**99**	**9**	**8**	**69**	**1**

The design of the optimal contract proceeds in two steps: First, the agent chooses a level of effort conditional on the contract. For Incentive Scheme 1, marginal analysis shows that the optimal level of effort for the agent is \$4: the marginal income to the agent from reducing a default is $0.5 \times 10 = \$5$, which exceeds the marginal cost of reducing defaults from three to two, but not from two to one. The top left graph in Figure 2 illustrates the agent's optimal decision graphically.

Top left panel shows the agent's decision to exert effort. Top right panel shows how the principal inverts the adjacent panel to infer effort from the incentive scheme. Bottom left panel shows the principal's optimal choice of effort without uncertainty, and the bottom right panel shows the optimal choice of effort with uncertainty. See Section 4 for details.

Figure 2
Graphical Depiction of a Simple Principal-Agent Problem Based on Holmstrom and Milgrom (1987)



In the second stage of the solution, the principal chooses an incentive scheme, taking the agent's optimal response as given. In other words, the principal scans Panel A of Table 1. Then, corresponding to each incentive scheme, the principal can read off levels of effort, defaults, and payments to the agent and, as a result, profits. Analysis of Panel A illustrates a central point of contract theory: despite being unable to observe effort directly, the principal can deduce the effort level from his understanding of the agent's optimal decision problem. The top right panel of Figure 2 makes the point that the principal can, in a sense, invert the agent's decision problem and, in doing so, choose the level of unobservable effort by choosing the incentive scheme.

Which is the optimal scheme? It is displayed in Panel B of Table 1. For example, for Incentive Scheme 3, the agent will put in maximal effort leading to zero defaults and a profit of $7 to the principal. Incentive Scheme 1 will generate much less effort, but the corresponding payment to the agent is smaller, so the profit stays the same at $7. Incentive Scheme 2 yields maximal profits with more effort than Scheme 1 but lower payments than Scheme 3.

What are the welfare implications? Incentive Scheme 2 is privately optimal, but is it socially optimal? Yes. The lower left panel of Figure 2 illustrates the principal's choice graphically. The principal's choice of one default maximizes the total social surplus and is thus socially optimal. Adam Smith's "invisible hand" at work again!

If the private outcome is socially optimal, then why did I say in the introduction that the First Welfare Theorem fails in the presence of asymmetric information? The reason is that there is no meaningful asymmetric information problem here. The principal cannot observe effort, but he can observe output, which is perfectly correlated with effort. To see the failure of the First Welfare Theorem, we need to introduce asymmetric information, and we do that by supposing that there is some random variation in the number of defaults. If the agent expends $4 in effort, the expected number of defaults equals two. But suppose, actually, with 50 percent probability, there will be one default and a 50 percent probability that there will be three.

Risky payoffs to the agent change the profit-maximizing contract. The optimal contract from Panel B now delivers the agent a lottery paying $72 and $90 with equal probabilities. For a risk-averse agent, the lottery is worth less than a certain payment of $81, meaning that

the agent's utility now falls short of $69 and—recall from earlier—we assumed that the agent had an outside option paying $69, so now the agent rejects the contract and refuses to work.

Assume, specifically, that the agent has negative exponential utility with an absolute risk aversion coefficient of 0.69, which implies that Incentive Scheme 2 yields a utility of $61, $8 less than it did with certainty. Similar analysis of Incentive Scheme 1 shows that the addition of uncertainty also lowers the utility of the contract to the agent. However, because risk retention is lower, the sensitivity of the payout is smaller (Incentive Scheme 1 pays out $68 and $78 with equal probabilities, as opposed to $73 with certainty), and so risk reduces utility from $69 to $65, that is, by half as much as for Incentive Scheme 2.

By raising the base payments, the principal can induce the agent to come back to work, and Panel C of Table 1 shows the uncertainty-adjusted incentive schemes. Panel C shows that uncertainty has inverted the ranking of the two incentive schemes: Incentive Scheme 1 is now profit maximizing. What changed? The key point here is that the addition of risk disproportionately affects Incentive Scheme 2 because of the higher level of risk retention.

The bottom right panel of Figure 2 illustrates the effects of adding uncertainty. The dashed line shows the marginal cost of different levels of effort and defaults. Without uncertainty, the principal simply had to compensate the agent for his effort. But now—to elicit a higher level of effort—the principal must increase risk retention, which, in turn, increases risk for which the principal must now compensate the agent. In other words, there are two components to the marginal cost of effort: the direct cost of compensating the agent for his time and the indirect cost of eliciting effort.

What about welfare? The failure of the First Welfare Theorem occurs here because of the indirect cost of eliciting effort. An all-knowing, all-seeing social planner would not need to use risk retention to get the agent to work and would choose the higher level of effort because the marginal benefit of reduced defaults exceeds the marginal cost of additional effort by the agent. Compared with that benchmark, a deadweight loss is represented by the shaded area under the marginal benefit line.

But is an all-knowing, all-seeing social planner the right benchmark? For such situations, economists have defined an alternate

welfare concept, called "constrained Pareto optimality," that limits the planner to the same information set as the one the principal has, meaning that the planner has to choose from the same incentive schemes as the principal. In designing the Dodd-Frank risk retention requirement, Congress implicitly acknowledged the idea of constrained optimality by imposing an incentive scheme and not a level of effort.

Does the risk retention requirement increase welfare? Suppose, in the model, Congress decided that it wanted first-best levels of effort and, as a result, imposed Incentive Scheme 2 on the principal. The bottom right panel of Figure 2 shows that such a rule would create a deadweight loss. Thus, the policy would eliminate the deadweight loss of reduced effort—the shaded area under the marginal benefit curve—but the cost of eliminating that deadweight loss would be the area between the two marginal cost curves, which, by construction, exceeds the deadweight loss from reduced effort; the overall deadweight loss is the shaded triangle above the marginal benefit line.

Before we continue, it is important to address three questions:

1. By reducing defaults, wouldn't a requirement of risk retention have attenuated the crisis? Yes and no. In theory, mandatory risk retention should have had no effect on the investors' massive losses, which caused the financial crisis. That statement may sound surprising, as we have shown that higher risk retention reduces defaults, but the problem here is that investors lose money not when more borrowers default, but when more borrowers default *than expected*. With more risk retention, investors would have expected more effort and fewer defaults and so, in the model, the losses would have been exactly the same.

At the same time, in the model, requiring risk retention does lead to fewer defaults. In a sense, one can think about retention as a tax on defaults: it causes a deadweight loss, but it does reduce defaults. In Section 5, we return to this question.

2. Isn't there empirical evidence showing that securitization caused lenders to expend less effort and thus contribute to the crisis? Many researchers cite a paper by Keys et al. (2010) as evidence that securitization led to lower levels of effort in underwriting, which, in turn, caused the crisis. Keys et al. purport to show that when lenders knew that there was a higher likelihood that they would sell the loan in a security, they did a worse job screening, thereby

leading to higher default rates.[3] The problem with interpreting Keys et al. as evidence in favor of the Dodd-Frank risk retention requirement is that their findings are completely consistent with the model described above. They show that more risk retention leads to more effort. Furthermore, in the model, more risk retention leads to more effort, as shown in Figure 2, but risk retention still reduces welfare. In other words, all Keys et al. do is confirm that the top left panel of Figure 2 is an accurate description of the world.

If we are to illustrate the point, suppose policymakers were considering an interest rate subsidy for manufacturers as a way of increasing manufacturing investment. What Keys et al. do is essentially equivalent to showing that lower interest rates lead to more investment, but it would obviously be wrong to conclude that manufacturing investment was suboptimal or that an interest subsidy would be welfare improving.

3. Doesn't the model imply that some risk retention is optimal? In the model, the optimal incentive scheme involves 50 percent risk retention, far in excess of what Dodd-Frank requires. Indeed, in the Holmstrom-Milgrom model (1987), the optimal scheme *always* involves some risk retention, so one might conclude that whether 5 percent is right or not, it is still better than the 0 percent that prevailed in many securitization deals in 2005. Doesn't that mean that policy should at least force lenders to retain *some* risk always, if not exactly 5 percent? No. The purpose of the model is to illustrate how the market determines the optimal level of risk sharing. The fact that firms in the real world did not exactly conform to the model means there is something wrong with the model, not that there is an opportunity for government to tell firms what to do.

As an illustration of the point, consider portfolio choice theory. All standard models of portfolio choice imply that the optimal allocation to stocks is greater than zero. But the data show that many households hold no equities. One might argue that, on the basis of the model, the government should enforce a minimum 5 percent allocation to stocks. But, instead, economists have tried to come up with economic explanations for why so few households hold stock.

[3] Bubb and Kaufman (2009) argue that Keys et al. (2010) misunderstood the institutional evidence and that the patterns they observed reflect underwriting rules having nothing to do with securitization. Nonetheless, for pedagogic purposes, we will assume that the interpretation by Keys et al. is correct.

4.1 Moral Hazard and Government Policy

The fact that the risk retention requirement enhances incentives but that a government policy to increase risk retention reduces welfare exposes a tension in economic thought. On the one hand, standard economic theory stresses the role of incentives. Adam Smith, the father of the "invisible hand," wrote:

> It is the interest of every man to live as much at his ease as he can; and if his emoluments are to be precisely the same, whether he does, or does not perform some very laborious duty, it is certainly his interest . . . either to neglect it altogether, or, . . . to perform it in [a] careless and slovenly a manner. (Smith [1776] 1904, p. 760)

On the other hand, Milton Friedman, Smith's modern disciple, makes a similar argument, laying out a simple theory:

> When you spend, you may spend your own money or someone else's; and you may spend for the benefit of yourself or someone else. Combining these two pairs of alternatives gives four possibilities summarized in the following simple table: (Friedman and Friedman 1980, p. 116)

	On Whom Spent	
Whose Money	**You**	**Someone Else**
Yours	I	II
Someone Else's	III	IV

Friedman then focuses on Category IV:

> Category IV refers to your spending someone else's money on still another person. You are paying for someone else's lunch out of an expense account. You have little incentive either to economize or to try to get your guest the lunch he will value most highly. (p. 117)

The tension here is that both Smith and Friedman were dedicated opponents of government intervention in markets and, I am quite certain, would have opposed the Dodd-Frank risk retention provision. But how could they oppose a policy that moves economic activity from Category IV of the table above to Category I? The key point of the

principal-agent model is that the principal knows that more risk retention leads to more effort but chooses less risk retention. In Friedman's example, the employer knows that workers in Category I work harder than do workers in Category IV; in a market economy, the employer is free to choose Category I if he wants to do so. The fact that he chooses Category IV despite the availability of Category I means that, for some reason, he believes that the benefits of Category IV versus Category I outweigh the costs.

Another way to see that point is to understand how the Dodd-Frank requirement affects the different market participants. Proponents view it as a limit on the behavior of the agent: the law, as written, circumscribes the behavior of the agent. But in practice, the law actually restricts the choice set of the principal: the law blocks an investor who wants to take on all the credit risk in a transaction from his desired choice. The implicit presumption of the framers of Dodd-Frank was that no reasonable person would want to invest in a security in which the issuer held no risk retention. But we turn to the central principle established by Smith and described by Friedman as follows: "Adam Smith's key insight was that both parties to an exchange can benefit and that, so long as cooperation is strictly voluntary, no exchange will take place unless both parties do benefit" (Friedman and Friedman 1980, p. 1).

4.2 More General Models

In the example in Table 1, government policy cannot increase welfare, but the result is substantially more general than that. The example is, as mentioned already, a special case of Holmstrom and Milgrom (1987), and the result extends to that model. But, in fact, Prescott and Townsend (1984) show that in any moral hazard model with a single consumption good, equilibrium is constrained to be efficient so the risk retention requirement always reduces welfare. In adverse selection models, Bisin and Gottardi (2006), however, show that equilibrium is typically constrained to be inefficient. In other words, government could increase welfare by restricting the space of available contracts. However, the general problem in adverse selection models is that there is too little risk sharing, not too much.

Consider the leading example, health insurance. Rothschild and Stiglitz (1976) consider a world where risk-averse individuals have private information about how healthy they are and show that the First Welfare Theorem fails because more healthy individuals fail to

insure fully—in other words, they retain some of their health risk. The solution, according to Rothschild and Stiglitz, and as implemented in the 2010 Patient Protection and Affordable Care Act, is to *force* individuals to buy insurance—to *ban* risk retention. In other words, if economists really believed that adverse selection was a problem in mortgage underwriting, the solution would not be to force lenders to retain bad loans, it would be to force them to securitize good loans!

In models with multiple goods, the situation becomes more complex. Changes in policy have general equilibrium effects on relative prices, and pretty much anything can happen. Greenwald and Stiglitz (1986), in a celebrated paper, show that, generically, governments can change the contract space in such a way as to increase welfare. The issue with asymmetric information is that agent behavior is circumscribed by a set of incentive compatibility constraints. As Greenwald and Stiglitz show, changes in relative prices can relax those constraints and thus lead to a welfare improvement.

If we are to illustrate why the existence of additional constraints allows government policy to improve welfare, consider some real-world examples. Suppose that lenders constrain households to borrow up to a specific multiple of income when they are buying homes. For households with an upward-sloping income profile, Gerardi, Rosen, and Willen (2010) show that such a constraint prevents smoothing of housing consumption over the life cycle. A government policy to drive down house prices relaxes the borrower's constraint and allows him to buy a bigger house and to better smooth consumption over the life cycle. An alternative example is that existing homeowners who want to move face a down payment constraint, as in Stein (1995). A government policy to increase house prices would relax the down payment constraint and allow better matching of households with homes.

In the Dodd-Frank Act, Congress points explicitly to the possibility that risk retention has broad effects. In Section 946, Congress asks for a "Study on the Macroeconomic Effects of Risk Retention Requirements." Specifically, Congress proposes "an analysis of the effects of risk retention on real estate asset price bubbles, including a retrospective estimate of what fraction of real estate losses may have been averted had such requirements been in force in recent years."

The link between risk retention and asset price bubbles is, at best, purely speculative. Economists have few good models of how bubbles form and no models that link low levels of risk retention to

bubbles. Indeed, the macroeconomic effects, almost by definition, are somewhat limited. To understand why, remember that the upper right panel of Figure 2 shows that principals can infer how much effort agents are putting in, meaning that, in the models at least, if investors invest in mortgage-backed securities with low levels of risk retention, they do so knowing that many borrowers will default. Therefore, they will pay a correspondingly low price.

Ultimately, pointing to the macroeconomic benefits of risk retention is also somewhat disingenuous. The main appeal of risk retention is its simplicity, which is displayed in the top left panel of Figure 2. More risk retention leads to more effort; that conclusion, at least, is settled in economics. As we have explained here, that finding alone, unfortunately, does not justify mandatory risk retention as a policy.

4.3 Irrationality

One potential justification for risk retention is that investors did not understand that there was a relationship between risk retention and effort. To see why investor misunderstanding could lead to an opportunity for government policy, imagine that, for example, the market outcome was Incentive Scheme 1 in Panel C of Table 1. But suppose investors believed that lenders misunderstood the incentives and believed that lenders were putting in $12 of effort, whereas the lender's optimal response to Incentive Scheme 1 was to exert only $4 of effort. Investors would then be shocked when twice as many defaults occurred as they expected. Such a result is broadly consistent with what happened in the crisis. If government policy forced lenders to use Incentive Scheme 2, then policy would bring investor beliefs into line with reality and their expectations of default into line with outcomes, potentially avoiding the financial crisis.

As a theory of the crisis, however, the idea that investors did not understand the incentives of lenders is problematic. As Foote, Gerardi, and Willen (2012) and Richardson, Ronen, and Subrahmanyam (2010) show, most of the firms with the greatest exposure to subprime risk were underwriters and securitizers of subprime mortgages. It seems implausible that Bear Stearns' executives would not have understood the link between effort and risk retention.

More broadly, investors based their beliefs about the performance of securitized mortgages on the historical performance of securitized mortgages. If no risk retention means no effort, then loans made

with no risk retention will perform badly and investors buying loans with no risk retention will pay accordingly. So it is hard to see how investors could have formed incorrect beliefs about the relationship between effort and retention.

5. RENEGOTIATION

During the crisis, many commentators lamented the unwillingness of lenders to renegotiate or "modify" mortgages. The logic was as follows: Suppose the borrower owes amount M^*, the house is worth $P < M^*$, and the lender will recover $(1 - \lambda)P$ from a foreclosure. If the lender sets the loan balance to $M' = (1 - \lambda)P$, the lender will be no worse off, and since $M' < P$, the borrower now has positive equity, can sell the property if needed, and has an incentive to keep making payments. Critics of the lending industry wondered why there were any foreclosures at all.

Throughout the crisis, there have been vigorous calls for executive action and legislation to force lenders to modify mortgages. For example, in a recent opinion piece, Martha Coakley and Eric Schneiderman (2013), attorneys general of Massachusetts and New York, respectively, wrote:

> Mortgage modification, including significant principal reduction for underwater mortgages, can actually increase the lifetime value of a mortgage by reducing the likelihood of default. It is far more profitable for any financial institution to hold a portfolio of performing $200,000 mortgages that keep families in their homes than a portfolio of nonperforming $250,000 mortgages headed toward default.

The most popular explanation for why lenders modified so few mortgages was institutional frictions particularly related to securitization. Since the entity making the decision about renegotiation—the servicer of the loan—did not actually own the loan, it did not stand to gain from modification and so generally opted against it. Subsequently, critics blamed the shortage of renegotiation on the intransigence of Edward DeMarco, the acting director of the Federal Housing Finance Agency, who had blocked principal reduction as a tool for the institutions he regulated, Fannie Mae and Freddie Mac.

One result is that there have been major policy changes with respect to delinquent loans. In the short run, the administration

implemented in 2009 the Home Affordable Modification Program, which provided subsidies to servicers with the goal of overcoming institutional frictions. But in addition to the emergency measures, policymakers have also made permanent changes to the relationship between borrower and lender. As part of the 2012 National Mortgage Settlement (NMS) with the state attorneys general, servicers agreed to a set of standards that gave borrowers substantial rights in the loan modification process. Although, in theory, no one has challenged the idea that the lender should maximize profits when conducting loss mitigation on delinquent loans, the NMS and other legal actions like the Multi-Agency Consent Decree and the California Homeowners Bill of Rights have established substantial rights for the borrower in the loss mitigation process.

However, as we will now discuss, it is not clear that the institutional friction theory of why modifications are rare is the correct one. As Adelino, Gerardi, and Willen (2013) show, the frictions could, at most, explain only a small part of the unwillingness of lenders to renegotiate loans. Figure 3, which is from that paper, shows that portfolio lenders who faced neither the frictions of private label securitization nor the strictures of Federal Housing Finance Agency regulations were not more likely to renegotiate mortgages. In short, securitization cannot explain why lenders failed to modify most mortgages.

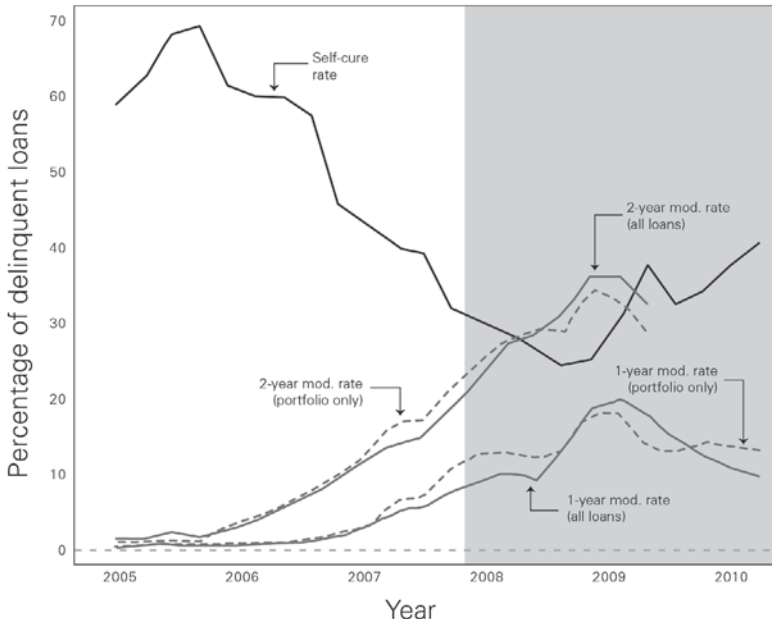
Why do lenders renegotiate so few mortgages? The economics of asymmetric information here provides a plausible explanation. To see why, return to the example at the beginning of this section. Now suppose that for each borrower, there is some amount V_i that he is willing to repay. Suppose that a continuum of borrowers is uniformly distributed along the interval $[0\ a/b]$. Suppose that the lender, instead of modifying the loan to $M' = (1 - \lambda)P$, sets the balance at $M > M'$ and forecloses on any borrower unwilling to repay M. Now,

$$Q = \frac{a - M}{a} \cdot \left(\frac{a}{b}\right) = \frac{a}{b} - \left(\frac{1}{b}\right) M$$

borrowers are willing to pay the modified balance, but overall the lender collects $M > (1 - \lambda)P$ from the Q borrowers who are willing to pay M and still collects $(1 - \lambda)P$ from the borrowers on whom it forecloses, and so the lender is better off. It is easy to see that to choose the optimal number of modifications Q, the lender solves the problem:

$$(3) \quad \max_Q M \cdot Q - (1 - \lambda)P \cdot Q = \max_Q (a - bQ) \cdot Q - (1 - \lambda)P \cdot Q$$

Figure 3
Loan Modifications, 2005–2011
Modifications are measured over the period (one year or two years) following the month that a loan became 60 days delinquent. See Adelino, Gerardi, and Willen (2013) for a full discussion of the methodology.



Source: Adelino, Gerardi, and Willen (2013).
Note: "Portfolio" refers to loans held on banks' balance sheets. "All loans" refers to loans securitized by government-sponsored enterprises and loans securitized by private firms.

Equation (3) should look familiar as it is the optimization problem of a monopolist facing a linear demand curve. Optimal Q solves the first-order condition

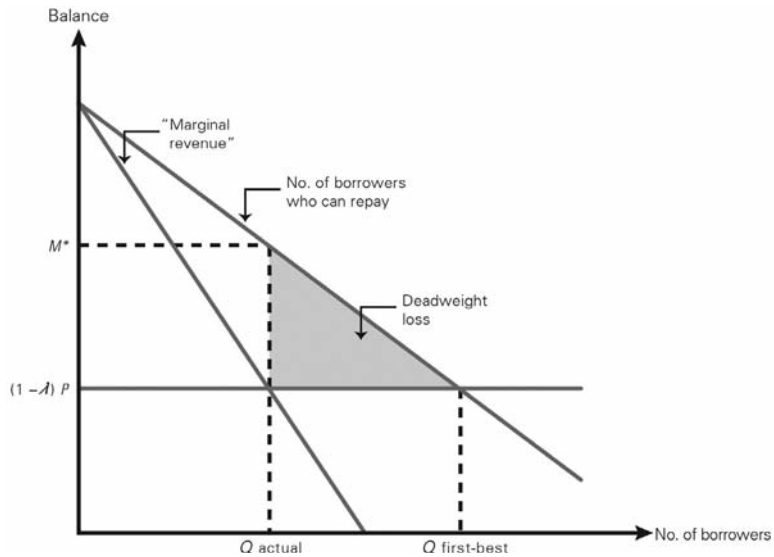
$$a - 2bQ = (1 - \lambda)P$$

where the left-hand side is the "marginal revenue" of an additional modification. Figure 4 illustrates the solution. The top line labeled "No. of Borrowers Who Can Repay" is Q as defined above, and we

can think of it as the demand for modifications. The "cost" of doing a modification is the revenue from the alternative, $(1 - \lambda)P$. The no-foreclosures solution, where the lender reduces principal to $(1 - \lambda)P$ for everyone, is the competitive solution. However, for the lender, the monopoly solution of setting the price equal to M^* obviously dominates the competitive solution.

Thinking about the modification problem as a monopoly pricing problem reconciles different views. On the one hand, critics of the industry were right that lender policy was leading to a large deadweight loss. In Figure 4, many borrowers were willing to pay more than the lender recovered from foreclosures. On the other hand, the view of the critics that lenders could increase profits by modifying more loans was wrong. The argument is precisely the same as saying

Figure 4
Loan Renegotiation
If borrowers have unobservably different willingness to repay their mortgages, then the lender's decision to modify is equivalent to a monopoly pricing problem.
See Section 5 for details.



that a hotel with empty rooms represents a deadweight loss but that, at the same time, cutting room rates to fill the rooms is not profit maximizing for the hotel.

What is the optimal policy here? In 2009, drawing on the logic that institutional frictions were the main reason lenders weren't modifying loans, the administration's Home Affordable Modification Program intervened by providing financial incentives to intermediaries to modify loans. The earlier analysis illustrates why relatively small financial incentives could not overcome the basic economics of loan renegotiation.

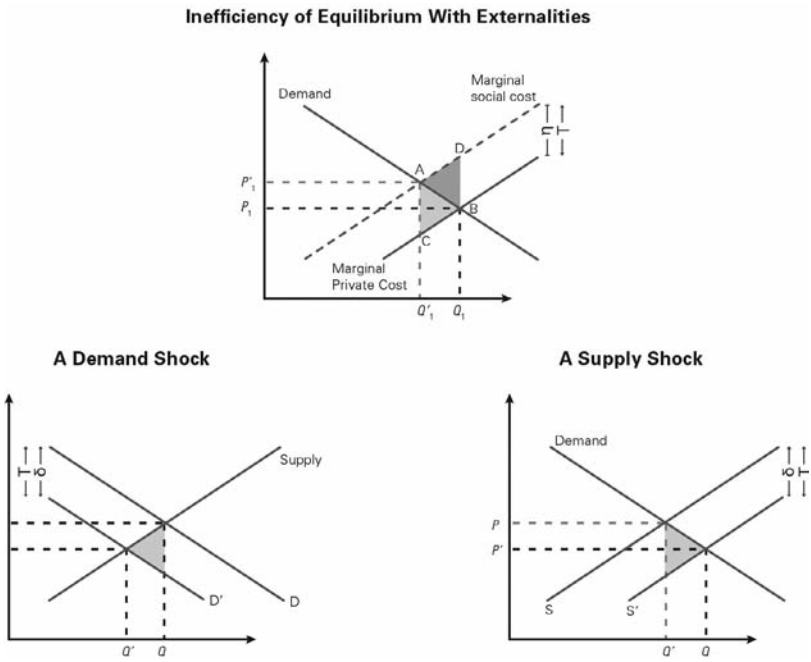
To prevent foreclosures, a government in Figure 4 has two options: First, it can force lenders to implement the competitive solution by requiring that lenders modify all loans down to $(1 - \lambda)P$. Although that option would inflict a large financial penalty on lenders, the total surplus would be substantially increased. Second, the government could force lenders to implement the competitive solution and use a tax to compensate them for lost profits. What should be clear, though, is that preventing foreclosures without inflicting losses on lenders or covering their losses is impossible.

Going forward, one could argue that a government policy that forced modifications would increase consumer surplus. The zero economic profit condition means that somewhere earlier in the process, lenders paid for the right to extract surplus from delinquent borrowers, and so the elimination of the deadweight loss will cost lenders nothing. Of course, in a sense, that surplus extraction was embedded in the price borrowers paid for the loans when they got them. So although the overall economic gain will be positive, some borrowers may complain that they would prefer to pay a lower rate up front and suffer the consequences later.

6. EXTERNALITIES

As mentioned in the Introduction, the presence of externalities in an economy invalidates the First Welfare Theorem. Consider a simple supply-and-demand model from Econ 101 with a downward-sloping demand curve and an upward-sloping supply curve. The top panel of Figure 5 illustrates the problem with externalities. Firms make decisions on the basis of the curve labeled "Marginal private cost," but production of the commodity inflicts an additional cost η on neighbors, which means that the "Marginal social cost" is higher.

Figure 5
The Effect of a Tax on the Sale of Property, Without and With Physical Externalities
If foreclosures cause physical externalities, then a tax can increase welfare (top panel). If foreclosures cause a pecuniary externality (e.g., an increase in the supply of properties on the market), then an offsetting tax is welfare reducing (bottom right panel). See Section 6 for details.



To understand why equilibrium is inefficient, imagine that we introduce a tax $\tau = \eta$ on producers. The tax reduces welfare by the triangle *ABC*, the usual deadweight loss resulting from the fact that consumer surplus and producer surplus fall by more than the tax revenue generated. But because production has fallen from Q to Q', the neighborhood costs fall by the quadrilateral *ADBC*. Subtracting the deadweight loss from the reduction in neighborhood costs yields the net benefit of the tax, the triangle *ABD*. Intuitively, the triangle

is a loss to society that results from the fact that the costs of the foreclosure to neighbors exceed the consumer and producer surpluses generated by the production. The first Q' of output still generates externalities, but from the standpoint of society as a whole, the tax revenue exactly offsets the costs. The fact that the government can increase welfare with a tax illustrates that the First Welfare Theorem cannot hold.

What are the externalities in the mortgage market, and how could government policies offset them to improve welfare? To answer that question, we need to distinguish between two types of externalities. The first type, depicted in the top panel of Figure 5, is known as technological or physical externalities, and they occur when something an individual or a firm does directly enters into the utility or production function of another individual or firm, respectively, in the economy. Pollution is the classic example, but for our purposes the more natural example is foreclosures. Thus, many have argued, foreclosures led to neglect of properties, which, in turn, inflicts damage on neighboring properties.

The second type of externality, referred to by Viner (1932) and elucidated by Scitovsky (1954) as pecuniary externality, is more subtle and works through market mechanisms. An example of a pecuniary externality occurs when I list my house for sale, thereby making it somewhat more difficult for someone selling a close substitute. Obviously, I don't take that into account when I list the property, just as I don't take into account the effect on my neighbors if I play loud music late at night or do compression tests on my motorcycle on a peaceful Saturday afternoon. Doesn't that provide another opportunity for government intervention? As we now explain, in our standard undergraduate models, the First Welfare Theorem still holds in the presence of pecuniary externalities: government cannot make everyone better off. In richer models, as we discuss next, pecuniary externalities do allow for welfare-improving government policies.

To see why government cannot help, consider first a demand shock, as depicted in the lower left panel of Figure 5. The left panel shows what happens if prices fall because of a demand shock δ. Suppose the government could introduce a subsidy $\tau = \delta$ that would restore the previous equilibrium prices and quantities. The figure illustrates that such a policy would reduce welfare, thereby generating a deadweight loss, shaded in the figure, which results from the fact that the subsidy leads transactions to occur between sellers who value the

properties less than the buyers do. Now, turning to the right panel of Figure 5, suppose a lender forecloses on a property and lists it for sale, shifting the supply curve by amount δ down and to the right and lowering prices and increasing the level of sales in the market. Suppose the government intervenes and introduces a tax $\tau = \delta$ on property sales, which exactly offsets the shift in the supply curve and restores prices to the previous level. Isn't that a good thing? No. The shaded triangle shows that there is deadweight loss to the economy: there are potential buyers who value properties more than potential sellers do, but because of the tax, those trades don't take place.

7. CONCLUSIONS

In the end, I believe that externalities have to be at the heart of any justification of the slate of policies discussed in this paper. Let me focus on risk retention because, in a sense, I view it as the most egregious error. As I have shown, risk retention does not solve the asymmetric information problem or "improve the securitization process." Instead, it simply makes lenders more cautious and thus reduces the number of defaults, but at the cost of a deadweight loss to the investors and lenders. If one believes that foreclosure externalities are significant, then one might view risk retention as a sort of Pigouvian tax[4] in which the social benefits of eliminating externalities make up for the deadweight loss.

But in practice, I am skeptical that risk retention would even achieve the goal of reduced defaults. As Foote, Gerardi, and Willen (2012) argue, the financial crisis resulted from the fact that most of the key financial intermediaries had too much exposure to residential real estate—in other words, they had *too much* skin in the game.

If externalities are the justification for risk retention, then one might well ponder a more direct approach: default taxes. In other words, risk retention is a roundabout way of preventing default that would have been ineffective anyway. A default tax targets the externality precisely and has the added benefit of generating revenue for the government at precisely the time when it is needed.

Academics and journalists are fond of saying that the crisis occurred because of "misaligned incentives." What they mean by

[4] A Pigouvian tax is a tax on negative externalities, which are effects that are harmful to another person or group.

that verbiage is never exactly clear. In economics, misaligned incentives lead to a breakdown in trade: a rating agency with misaligned incentives is not worth anything to issuers or to the owners of the rating agency. I think what people have in mind is that the incentives of market participants are not aligned with the goals of society; in other words, that the invisible hand has failed. Default externalities are a perfect example. Because they do not incorporate the costs of foreclosures on neighbors, market prices do not give lenders proper incentives to avoid defaults, and the number of defaults in equilibrium is suboptimally high. Yes, incentives are misaligned, but no, the incentive misalignment has nothing to do with the structure of private financial contracts. Distorting financial contracts, as key provisions of Dodd-Frank do, can help only by accident.

REFERENCES

Adelino, Manuel, Kristopher Gerardi, and Paul S. Willen. 2013. "Why Don't Lenders Renegotiate More Home Mortgages? Redefaults, Self-Cures and Securitization." *Journal of Monetary Economics* 60 (7): 835–53. http://dx.doi.org/10.1016/j.jmoneco.2013.08.002.

Akerlof, George A. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84 (3): 488–500.

Bisin, Alberto, and Piero Gottardi. 2006. "Efficient Competitive Equilibria with Adverse Selection." *Journal of Political Economy* 114 (3): 485–516.

Bolton, Patrick, Xavier Freixas, and Joel Shapiro. 2012. "The Credit Ratings Game." *Journal of Finance* 67 (1): 85–111.

Bubb, Ryan, and Alex Kaufman. 2009. "Securitization and Moral Hazard: Evidence from a Lender Cutoff Rule." Technical Report no. 09-5, Public Policy Discussion Papers, Federal Reserve Bank of Boston.

Coakley, Martha, and Eric Schneiderman. 2013. "Housing Agency Needs New Leader." *Politico*, April 3.

Foote, Christopher L., Kristopher S. Gerardi, and Paul S. Willen. 2012. "Why Did So Many People Make So Many *Ex Post* Bad Decisions? The Causes of the Foreclosure Crisis." NBER Working Paper no. 18082, National Bureau of Economic Research, Cambridge, MA.

Friedman, Milton, and Rose D. Friedman. 1980. *Free to Choose: A Personal Statement*. Orlando, FL: Harcourt.

Geanakoplos, John, and Herakles M. Polemarchakis. 1986. "Existence, Regularity and Constrained Suboptimality of Competitive Allocations When the Asset Market Is Incomplete." In *Uncertainty, Information and Communication: Essays in Honor of Kenneth J. Arrow*, vol. 3, ed. Walter P. Heller, Ross M. Starr, and David A. Starren, pp. 65–96. Cambridge: Cambridge University Press.

Gerardi, Kristopher S., Harvey S. Rosen, and Paul S. Willen. 2010. "The Impact of Deregulation and Financial Innovation on Consumers: The Case of the Mortgage Market." *Journal of Finance* 65 (1): 333–60.

Greenwald, Bruce C., and Joseph E. Stiglitz. 1986. "Externalities in Economies with Imperfect Information and Incomplete Markets." *Quarterly Journal of Economics* 101 (2): 229–64.

Gul, Faruk, and Wolfgang Pesendorfer. 2001. "Temptation and Self-Control." *Econometrica* 69 (6): 1403–35.

Holmstrom, Bengt, and Paul Milgrom. 1987. "Aggregation and Linearity in the Provision of Intertemporal Incentives." *Econometrica* 55 (2): 303–28.

Keys, Benjamin J., Tanmoy Mukherjee, Amit Seru, and Vikrant Vig. 2010. "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *Quarterly Journal of Economics* 125 (1): 307–62.

Keys, Benjamin J., Tomasz Piskorski, Amit Seru, and Vikrant Vig. 2013. "Mortgage Financing in the Housing Boom and Bust." In *Housing and the Financial Crisis*, ed. Edward Glaeser and Todd Sinai, pp. 143–204. Chicago: University of Chicago Press.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112 (2): 443–78.

Prescott, Edward C., and Robert M. Townsend. 1984. "Pareto Optima and Competitive Equilibria with Adverse Selection and Moral Hazard." *Econometrica* 52 (1): 21–45.

Richardson, Matthew P., Joshua Ronen, and Marti Subrahmanyam. 2010. "Securitization Reform." In *Regulating Wall Street: The Dodd-Frank Act and the New Architecture of Global Finance*, ed. Viral V. Acharya, Thomas F. Cooley, Matthew P. Richardson, and Ingo Walter. Hoboken, NJ: John Wiley.

Rothschild, Michael, and Joseph Stiglitz. 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Quarterly Journal of Economics* 90 (4): 629–49.

Scitovsky, Tibor. 1954. "Two Concepts of External Economies." *Journal of Political Economy* 62 (2): 143–51.

Smith, Adam. (1776) 1904. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Methuen & Co., Ltd.

Stein, Jeremy C. 1995. "Prices and Trading Volume in the Housing Market: A Model with Down-Payment Effects." *Quarterly Journal of Economics* 110 (2): 379–406.

Viner, Jacob. 1932. "Cost Curves and Supply Curves." Reprinted in *The Long View and the Short: Studies in Economic Theory and Policy*, ed. Jacob Viner. Glencoe, IL: Free Press.

Comment

Neng Wang

The recent financial crisis has had a profound effect on the macroeconomy. Various policy interventions in response to the crisis have been proposed and implemented. Paul Willen challenges the soundness of some of those policies by using modern economic theory (e.g., classical welfare analysis and modern agency theory). Specifically, he provides a critical assessment of the following four proposed regulations and policies:

- the ability-to-repay requirement in mortgage underwriting
- reform of rating agency compensation
- risk retention in securitization
- mandatory loan renegotiation

Willen's main argument against those policies is quite straightforward: Standard economic theory leaves little room, if any, for government intervention. Rational economic agents living in the real world and facing complicated frictions (e.g., informational asymmetry and various forms of moral hazards) already behave in their own interests by optimally choosing their actions and designing constrained efficient contracts. It is thus very hard to improve the equilibrium outcome via government interventions. Importantly, it is insufficient to justify government intervention simply because the observed economic outcomes (e.g., the costly foreclosure process) appear highly undesirable. One needs to provide an argument on how government intervention can create additional net value beyond what the private market and contractual agreements can deliver.

Willen nicely frames the debate on the validity of those policies by centering the discussions on the economics-based reasoning (Economics 101 in action). Although providing a natural benchmark for us to think critically about those policies, he does not rule out the

Neng Wang is the Chong Khoon Lin Professor of Real Estate and Finance at Columbia Business School and a research associate with the National Bureau of Economic Research.

possibility that the policies may be justifiable in a richer economic setting with more important real-world frictions. For instance, he writes, "I believe that externalities have to be at the heart of any justification of the [crisis consensus] slate of policies."

Understandably for expositional purposes, the standard models that he uses to illustrate his key argument summarized earlier are stylized and may not incorporate some important real-world frictions. For example, externalities, bubbles, institutional constraints, general equilibrium considerations, and some forms of bounded rationality, irrationality, or both (at least for some agents in the economy) may provide some justifications for some forms of government intervention.

I will now briefly comment on Willen's critique of the policies. As an example, consider the ability-to-repay requirement for mortgage underwriting. The significant deterioration of mortgage underwriting standards during the precrisis period is often viewed as a cause of the recent financial crisis. With the objective of tightening the underwriting standard, the 2010 Dodd-Frank Act requires that "no creditor may make a residential mortgage loan unless the creditor makes a reasonable and good faith determination based on verified and documented information that, at the time the loan is consummated, the consumer has a reasonable ability to repay the loan, according to its terms, and all applicable taxes, insurance (including mortgage guarantee insurance), and assessments."

An economist may ask: "If imposing the ability-to-repay requirement is a good idea, why don't private parties choose to do so in their own interests? If verifying and documenting information create net surplus, why don't the borrower and lender get together and find a way to implement that? What additional value will this ability-to-repay regulation create, if any?" In competitive markets, standard economic theory will predict that contracts between the borrower and lender will be optimally chosen so that the lender's value and the borrower's value will be on the Pareto frontier, and hence the government (or any other third party) cannot improve the contractual agreement between the private parties. Indeed, imposing certain requirements, such as ability to repay or a minimal level of the lender's verification effort, effectively introduces an additional constraint and distorts the optimal contracting agreement between the lender and borrower, reducing the total surplus.

That is a classic and powerful argument used in various economic applications. Unless the government has information that private

parties lack (which seems unrealistic), or can resolve market failures (e.g., externality) that private agents have no incentives to address, or can achieve a more efficient allocation at a lower cost than private parties do (subject to the incentive-compatibility conditions from the private sector), there is little room for government intervention, regardless of how seemingly desirable the policy may appear.

On a related point, the government should not simply limit "exotic" mortgages (e.g., adjustable-rate mortgages with a teaser rate and a likely follow-up rate increase) just because the complicated mortgages are often perceived to be too risky for subprime borrowers and perhaps have been used by sophisticated lenders to take advantage of naive borrowers. Instead, using a state-of-the-art dynamic contracting framework, Piskorski and Tchistyi (2010) show that "exotic" mortgages (with complicated features, including adjustable rates, progressively increasing payments, and prepayment penalties) can benefit less creditworthy households in contrast to the popular view.

Having said that, one has to acknowledge that standard economic theory does make some strong assumptions that inevitably have strong policy implications (e.g., on mortgage lending practices). We need to think about the robustness of those policy implications once we allow for deviations from the standard assumptions. First, borrowers are assumed to be rational in standard economic models. There is much evidence at the micro level indicating that borrowers may be irrational, especially when facing very complicated financial products. An adjustable-rate mortgage with a teaser rate and various implicit or explicit state contingencies embedded in the contract are simply very difficult for many households to understand. Even economic theorists have only recently figured out the economic settings under which adjustable-rate mortgages are optimal contracts.

Additionally, some borrowers may have time-inconsistent preferences and have difficulties with making valuable commitments (Laibson 1997; Gul and Pesendorfer 2001). For borrowers with those behavioral biases, imposing some constraints (e.g., ability to repay) might help borrowers make better financial decisions (e.g., to mitigate time-inconsistency problems). Willen also notes the possibility of government policy in such a world, but he thinks that behavioral assumptions are unlikely to justify the ability to repay or other policies. The correct view is unclear to me at this moment. We need more economic models (e.g., contracting) where agents are subject

to bounded rationality and behavioral biases to further assess the implications of important behavioral biases on government policies.

Also, the general equilibrium implications of micro frictions (e.g., the one between borrower and lender) on the macroeconomy can be very different from those in the standard micro-agency models (not cast in general equilibrium). Why is that so? As an example, systemic risks (particularly important for the crisis-related policy debates) at least partly created by the increasingly sophisticated financial intermediation sector are not factored into any models, including optimal contracting models. Therefore, there may be additional social costs imposed by private parties on the macroeconomy, as private parties do not fully bear the costs on society—a form of externality, as Willen puts it. With the possibility of systemic risks, private contractual agreements in theory can be potentially enhanced via government intervention. Additionally, if private agents in the economy have the expectations that the government will bail them out in really bad times (e.g., the crisis period), the optimal contracting arrangement between private agents at the micro level will not be desirable at the aggregate level, as the government effectively is an agent (it has to make payments in crisis!) but is not involved in the contracting stage between private agents.

Another potentially important contributor to the recent crisis is the housing bubble. Does standard economic theory yield sensible policy implications in a world that may have housing bubbles or mispricing? How does the housing bubble influence the implications of policy interventions? Willen mentions the potential relevance of unrealistic house price appreciation on policy (e.g., the ability-to-repay requirement). What are the incentives for borrowers and lenders in a world with unrealistic house price appreciation? The incentive issues between borrower and lender become much more complicated with bubbles. How should the government behave in a setting with endogenous bubbles? Interestingly, in a bilateral contracting setting, Piskorski and Tchistyi (2011) show that high expectations of house price growth can help explain some controversial features of the recent subprime lending episode, such as loans with low initial rates but set to increase over time, but they show only the results in a partial equilibrium setting with exogenous price appreciation. The implications of policy interventions in general equilibrium with endogenous bubbles remain unexplored.

Willen has written a provocative paper that puts economic theory at the center of important policy debates. By using standard microeconomic

theory with rational agents and no externality, he shows that the room for effective government intervention to improve welfare is rather limited. Intuitively, economic agents have the incentives on their own to work out the optimal contractual agreements among themselves, and the government has no particular advantage to improve the private resource allocation and contract design. Willen's article serves as an excellent starting point for further constructive and deeper crisis-related policy debates. However, because some important frictions are left out of stylized partial-equilibrium models, we do not yet have conclusive answers on important crisis-related policy debates. We need richer general-equilibrium models with important frictions to evaluate those policies.

One frequently quoted argument in support of the government intervention in the recent crisis is the fear that, otherwise, the economy could have suffered much more (a counterfactual that we will never observe in reality). To evaluate various crisis-motivated policies, we need richer tractable, dynamic, and intuitive models that allow us to incorporate systemic risk and externality, bubbles, bounded rationality, and various important micro frictions (e.g., incentive issues, informational frictions, and institutional constraints) in order to better assess the consequences of various policies.

Policy debates eventually boil down to details. (I guess no one knows where the 5 percent risk retention in Dodd-Frank comes from.) What are the quantitative implications and economic significances of those policies? We can only try to answer those questions in quantitative tractable equilibrium models with important frictions embedded, as discussed earlier.

Recently, we have seen a fast-growing list of PhD job-market candidates doing promising financial crisis–related research that lies at the intersection between macroeconomics and finance. The best predictor for future exciting research is the junior economists' job market. I am optimistic that we will learn much about financial crisis and crisis-related policies in the next decade.

REFERENCES

Gul, Faruk, and Wolfgang Pesendorfer. 2001. "Temptation and Self-Control." *Econometrica* 69 (6): 1403–35.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112 (2): 443–78.

Piskorski, Tomasz, and Alexei Tchistyi. 2010. "Optimal Mortgage Design." *Review of Financial Studies* 23: 3098–3140.

———. 2011. "Stochastic House Appreciation and Optimal Mortgage Lending." *Review of Financial Studies* 24: 1407–46.

Comment

Parag Pathak

Paul Willen's wide-ranging and provocative paper offers a lot to consider. To summarize, he examines four policy responses to the recent housing market bubble and financial crisis, and he makes the following claim: simple economic models do not necessarily imply a role for government policy. "Simple" is an important word here, for reasons that I'll explain below.

Three of the four policy responses Willen examines are especially noteworthy because they're intended to harness private information that banks and other financial firms supposedly had, or should have had, going into the financial crisis that might have averted the crisis. These three policy responses are as follows:

- requirements for risk retention and securitization ("skin in the game")
- policies related to the "issuer pays" credit rating system
- policies concerning borrower "ability to repay"

Willen argues that, contrary to those policies, regulators are subject to the same kinds of information problems as market participants, and so we should be somewhat skeptical of whether any of those policies actually lead to Pareto improvements.

I really appreciate Willen's emphasis on first principles. It is my opinion that regulatory reforms are too often rushed in the wake of a crisis, ignoring simple yet important insights from the field of economics. There's a fair amount of discussion in his paper about first principles of economics—back to the insights of Milton Friedman and Adam Smith. That is where I want to begin my comments: I want to situate my remarks on my take of how basic welfare economics has evolved.

Willen begins his paper by discussing the First Welfare Theorem: the idea that competitive markets lead to Pareto-optimal outcomes so long as there are no market failures. For government intervention

to produce a Pareto improvement, some sort of market failure must exist—a failure like asymmetric information, externalities, or some sort of missing markets problem. Three of the four policy responses Willen considers are efforts to address a supposed asymmetric information problem. To his eye, the contracting party that supposedly is the beneficiary of the information asymmetry is the party that suffered the most harm from the suspect transaction in the run-up to the crisis, so it makes little sense to say that policy intended to right those asymmetries would produce a better outcome. To my eye, however, the details in those transactions are incredibly important, and it is essential to precisely model the contractual frictions he has in mind. It is not sufficient to appeal to general theorems in general equilibrium with adverse selection and moral hazard as Willen does, because those papers illustrate how rich commodity spaces are needed to restore efficiency, and they may not be realistic.

Basic welfare economics has also advanced to incorporate political economy considerations. Willen argues that this feature further undermines the rationale for interventions. The second way welfare economics has evolved is that it now considers the challenge posed by behavioral economics. Behavioral economists argue that people make mistakes, they may be misinformed, and they may have wrong beliefs. Those realizations, by themselves, do not provide sufficient rationale for intervention. It is not enough for proponents of policy intervention to recognize *ex post* that mistakes were made; Pareto-improving intervention requires that recognition *ex ante*. The burden is on the proponents to identify precisely what mistakes are looming. I am sympathetic to Willen on that matter, and the policy implications of that perspective can be as radical as he suggests. However, he should acknowledge the possibility that some interventions may make the market more operable or make it easier for market forces to play out their role.

Let's now consider more carefully some of the interventions that he discusses. Consider first the "skin in the game"—requiring banks to retain some ownership over the mortgage-related financial instruments they sell off. Specifically, a provision in the Dodd-Frank legislation says that securitized mortgage lenders are required to retain at least 5 percent of the credit risk for any security they issue that is backed by a mortgage. If there is securitization without skin in the game, Willen postulates that rational market participants will demand compensation for the supposed lack of mortgage

lenders' monitoring. He clearly describes the apparent tension between the idea that if mortgage lenders retain some of the risk, then they'll improve their monitoring of mortgages, and the possibility that the benefit of fewer defaults will be offset by higher mortgage costs, reducing public welfare.

He examines what happens when people have biased beliefs, overestimating underwriting efforts and underestimating defaults. He shows that it is possible for policies that force mortgage lenders to retain some risk to enhance public welfare by debiasing investors. That realization draws on work that Willen has undertaken in other papers, where he examines how firms that were the underwriters of securitizers of subprime mortgages would then leave themselves heavily exposed to those risky mortgages. That is, how could those firms have underestimated that risk so dramatically? I think that is a very challenging question, and it might undermine taking an Econ 101 view on the crisis.

Willen has been a leader among academics in trying to tease out whether those behavioral biases were actually taking place. In a paper with Kristopher Gerardi et al. (2008), he argues that many mortgage market participants understood that if house prices fell, then many borrowers would default. The paper includes Lehman Brothers' estimates in 2005 of defaults and losses under various scenarios, including a "financial meltdown" scenario. Lehman understood the risk modeling, but it apparently got the actual trajectory of house prices incorrect. That observation would support the idea that people systematically had the wrong impressions about future house prices. Another table in the paper illustrates how J. P. Morgan analysts continuously reappraised the likelihood of different types of house price changes and their effects. In 2006 and 2007, we see signs that the analysts' belief-updating in response to real-world news was slow, but that it improved in the fall of 2007 as price declines accelerated, as well as in December 2007 when there was some stabilization. The delayed reactions of even those sophisticated participants beg the question of what we can assume about market participants.

Now let's turn to policies involving financial rating firms. One premise underlying the Dodd-Frank legislation is that the inaccuracies of the financial ratings of unstructured financial products in the run-up to the financial crisis was due, at least in part, to a conflict of interest for the rating firms. Under the system that was then in

place, firms that created the financial products paid the rating firms for analyzing and grading the products. That practice supposedly created the conflict: the rating firms may not have wanted to deliver low grades to their customers. Dodd-Frank considers replacing that payment system with one that basically treats ratings as a public utility. Willen, however, points out a problem with the conflict-of-interest theory: the biggest losses in the crisis were experienced by the financial firms that issued the securities. That is, if a conflict of interest existed, then the financial firms paid money to be misled—a theory that does not make much sense. So again, the onus is on policy proponents to demonstrate that investors did not understand the conflict of interest.

On the subject of ratings, I believe that Coval, Jurek, and Stafford (2009) present compelling evidence that investors misperceived risks with ratings. Those authors argue that a fundamental principle of asset pricing is that assets that deliver returns in bad economic times—or, more formally, in states where the endowment is low—should be priced higher than other assets because the former deliver dollars when those dollars are most valuable. They draw a distinction between the structured finance products that were used to finance the housing boom—credit default obligations (CDOs) and the like—and catastrophe bonds. Catastrophe bonds pay off when a catastrophe occurs, and so they should have a relatively high price, whereas the housing products should have a lower price. However, the latter financial products were priced relatively high during the housing boom because they were priced based on their credit ratings alone. Coval, Jurek, and Stafford show empirically that if you look at the pricing of a lot of structured products, even though they should demand different risk premiums than, say, AAA-rated corporate debt, they did not actually command different risk premiums. The authors note that covariances are what is really important in asset pricing—something that was not well understood in the practitioner literature on the pricing of CDOs.

The issue here is that rating agencies professed CDOs and the like to be safe even though they were not, from an asset-pricing perspective. As Willen has noted, maybe the problem here is that investors rely too heavily on credit ratings. In particular, some market participants were restricted to having ownership of only highly *rated* securities, so the risks (according to Coval, Jurek, and Stafford) were

not properly understood and not properly priced by the markets. You can interpret that situation as systematic evidence that people did not understand and price the rating firm conflict of interest. Willen responds by saying that, sure, people did not anticipate those risks back in 2005 and 2006, but they won't make the same mistake going forward. That is a fair point, but it relies on not repeating the same mistakes, which seems like a strong assumption.

Let's now consider the ability-to-repay requirements. Did borrowers understand the mortgages that they took out? As Willen explains, the conventional wisdom is that they did not. That answer takes us beyond the world of the simple economic model, which assumes rational agents with perfect information and perfect foresight. As Willen explains, if the borrower has private information, a rational lender should simply verify income up to the point where it is in the lender's interest.

Woodward and Hall (2012) make the interesting argument that mortgage brokers would benefit greatly if they would "shop around" more vigorously for borrowers just as would-be borrowers can shop for mortgage lenders. That "shopping around" process would help expose would-be borrowers' private information. The authors go on to argue that a simplified environment where a broker receives all compensation from a lender rather than points and other fees would likely lead to better terms for the borrower as well as aid the lender. That argument brings us back to behavioral economics and the notion that borrowers may have self-control issues.

Woodward and Hall adopt an explicit behavioral point of view: "We are inclined to believe that simple admonitions, such as 'mortgage brokers are salesmen and the only way to get a good deal is to shop and bargain' and 'you are more likely to get a good deal if you shop for no-cost loans' are more likely to yield improvements than, for example, trying to teach borrowers enough financial economics to understand the tradeoff between cash and the interest rate" (p. 3271). I have a hard time squaring that idea, which I think is probably right, with the simple Economics 101 logic that Willen heavily relies on, because the inherent idea is that people do not actually process enough information and shop around enough. Woodward and Hall are basically saying that it does not even make sense to teach prospective borrowers basic financial economics because taking out a mortgage is a complex problem. Speaking more generally, I believe that notion

underlies many of the proposed policies advocated in response to the financial crisis.

Let's now turn to the final policy measure Willen examines in the paper, loan modification, which does not fit under the "asymmetric information" category. Foreclosures in large numbers appear to create a sort of negative externality on empty neighborhoods, with resulting harms in public safety, quality of life, housing market pricing, and public finance (see, e.g., Campbell, Giglio, and Pathak 2009). That externality provides justification for policy intervention, and one proposed intervention is policies to encourage loan modifications that would allow mortgage holders to stay in their homes but make more manageable payments.

The conventional wisdom is that there haven't been enough loan modifications in the wake of the financial crisis because of "institutional frictions" that Willen describes in his paper. He argues, in effect, that lenders have imperfect information about borrowers' willingness to pay, and that the lenders would be willing to modify if they had better information because the returns following a large number of modifications would be preferable to the returns following a large wave of foreclosures. To overcome the frictions, we could use some small financial incentives, which is the thinking behind the federal Home Affordable Modification Program. I think Willen is relatively spot-on in saying that, in effect, we want to change the profit maximization calculus of banks, and small financial incentives may not be enough.

In that setting, another possible policy response is to inflict losses on foreclosing lenders, reflecting the loss to public welfare from foreclosure externalities. So how do the foreclosure externalities compare with the costs—both private and public—now experienced by lenders? That question then leads to a second question: once policies are in place to encourage modification, will nondistressed borrowers then act strategically to force remodification, which could harm public welfare in the long run? There is now an emerging literature on that issue and the potential manipulation of eligibility criteria for mortgage modification. Guiso, Sapienza, and Zingales (2013) argue that borrowers generally will not act strategically because of bounded rationality or moral considerations. Mayer et al. (2011), on the other hand, looked at modifications made in the wake of Countrywide Financial's settlement with state attorneys general and

found strong evidence of strategic behavior: the delinquency rate of Countrywide's loans increased after the mortgage modification program was announced. My own view is that if we are going to follow Econ 101, then the potential for strategic behavior must be part of the discussion.

To conclude, I think it is hard to disagree with Willen's basic premise that simple economic models do not rationalize intervention, but the issue I raise is whether simple economic models are relevant for describing the financial crisis. I think there is a larger role for behavioral considerations. Willen is not opposed to that view, and we both agree that the arguments need to be articulated clearly. There is always the long-standing challenge of understanding whether market forces will lead people to learn and overcome their behavioral biases. I think the most important feature of Willen's paper and his other work is that it points to our need to better understand what went wrong and whether that was knowable *ex ante*.

REFERENCES

Campbell, John, Stefano Giglio, and Parag Pathak. 2009. "Forced Sales and House Prices." *American Economic Review* 101 (5): 2108–31.

Coval, Joshua D., Jakub Jurek, and Erik Stafford. 2009. "The Economics of Structured Finance." *Journal of Economic Perspectives* 23 (1): 3–25.

Gerardi, Kristopher, Andreas Lehnert, Shane M. Sherlund, and Paul S. Willen. 2008. "Making Sense of the Subprime Crisis." *Brookings Papers on Economic Activity* 2: 69–145.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2013. "The Determinants of Attitudes toward Strategic Default on Mortgages." *Journal of Finance* 68 (4): 1473–515.

Mayer, Christopher J., Edward Morrison, Tomasz Piskorski, and Arpit Gupta. 2011. "Mortgage Modification and Strategic Behavior: Evidence from a Legal Settlement with Countrywide." NBER Working Paper no. 17065, National Bureau of Economic Research, Cambridge, MA.

Woodward, Susan E., and Robert E. Hall. 2012. "Diagnosing Consumer Confusion and Sub-Optimal Shopping Effort: Theory and Mortgage-Market Evidence." *American Economic Review* 102 (7): 3249–76.

Cato Institute

Founded in 1977, the Cato Institute is a public policy research foundation dedicated to broadening the parameters of policy debate to allow consideration of more options that are consistent with the principles of limited government, individual liberty, and peace. To that end, the Institute strives to achieve greater involvement of the intelligent, concerned lay public in questions of policy and the proper role of government.

The Institute is named for Cato's Letters, libertarian pamphlets that were widely read in the American Colonies in the early 18th century and played a major role in laying the philosophical foundation for the American Revolution.

Despite the achievement of the nation's Founders, today virtually no aspect of life is free from government encroachment. A pervasive intolerance for individual rights is shown by government's arbitrary intrusions into private economic transactions and its disregard for civil liberties. And while freedom around the globe has notably increased in the past several decades, many countries have moved in the opposite direction, and most governments still do not respect or safeguard the wide range of civil and economic liberties.

To address those issues, the Cato Institute undertakes an extensive publications program on the complete spectrum of policy issues. Books, monographs, and shorter studies are commissioned to examine the federal budget, Social Security, regulation, military spending, international trade, and myriad other issues. Major policy conferences are held throughout the year, from which papers are published thrice yearly in the Cato Journal. The Institute also publishes the quarterly magazine Regulation.

In order to maintain its independence, the Cato Institute accepts no government funding. Contributions are received from foundations, corporations, and individuals, and other revenue is generated from the sale of publications. The Institute is a nonprofit, tax-exempt, educational foundation under Section 501(c)3 of the Internal Revenue Code.

CATO INSTITUTE
1000 Massachusetts Ave., N.W.
Washington, D.C. 20001
www.cato.org