Federal Reserve Bank of St. Louis

REVIEW

JANUARY/FEBRUARY 2009 VOLUME 91, NUMBER 1

Review is published six times per year by the Research Division of the Federal Reserve Bank of St. Louis and may be accessed through our website: **research.stlouisfed.org/publications/review**. All nonproprietary and nonconfidential data and programs for the articles written by Federal Reserve Bank of St. Louis staff and published in *Review* also are available to our readers on this website. These data and programs are also available through Inter-university Consortium for Political and Social Research (ICPSR) via their FTP site: www.icpsr.umich.edu/pra/index.html. Or contact the ICPSR at P.O. Box 1248, Ann Arbor, MI 48106-1248; 734-647-5000; **netmail@icpsr.umich.edu**.

Single-copy subscriptions are available free of charge. Send requests to: Federal Reserve Bank of St. Louis, Public Affairs Department, P.O. Box 442, St. Louis, MO 63166-0442, or call (314) 444-8809.

General data can be obtained through FRED (Federal Reserve Economic Data), a database providing U.S. economic and financial data and regional data for the Eighth Federal Reserve District. You may access FRED through our website: **research.stlouisfed.org/fred**.

ISSN 0014-9187

Three Funerals and a Wedding

James B. Bullard

This article is a modified and updated version of a speech presented at the Regional Economic Summit, Evansville, Indiana, November 20, 2008.

Federal Reserve Bank of St. Louis *Review*, January/February 2009, *91*(1), pp. 1-12.

The U.S. economy continues to face substantial turmoil. Financial markets are under unusual stress. Wall Street has been racked by seismic change. Uncertainty over the future prospects for the U.S. economy has caused consumers and businesses to pull back on discretionary consumption and investment spending. Doubts concerning the true value of complex securities continue to weigh heavily on financial markets worldwide. The still-uncertain fate of housing markets has kept the value of the underlying mortgage assets obscured.

The Federal Reserve has been active and innovative in responding to the evolving turmoil during 2008. In addition to deploying interest rate cuts, the Fed has implemented a series of new and unconventional tools. This innovation has intensified in response to evolving market events. There may be many more twists and turns in the policy response going forward.

I will discuss the challenges my Federal Reserve colleagues and I face as we strive to implement a policy that is designed to deliver low and stable inflation along with maximum sustainable employment. I will describe three funerals and a wedding—that is, three ideas about the U.S. economy that may be going to their final resting place and one idea that, once left for dead, may be taking on a new life. I will keep you in suspense about what ideas I have in mind.

As always, any views expressed here are my own and do not necessarily reflect the official views of other Federal Open Market Committee members.

THE FATE OF THE GREAT MODERATION

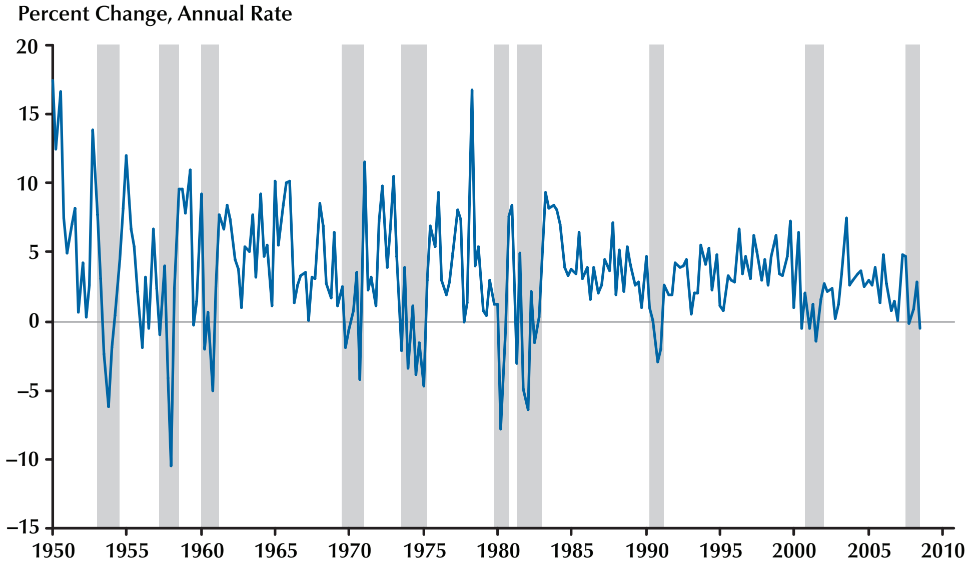
A common description of current events is that some cherished theories about the macroeconomy have been shattered. One idea is that the fabled resiliency of the U.S. economy over the past several decades is being called into question. Policymakers and academics alike have described the period since the mid-1980s as the Great Moderation, meaning that the volatility of the economy has been markedly lower during recent decades than it was in the earlier part of the postwar era, and certainly much less than during the interwar period during the 1920s and 1930s. Now, that moderation and resiliency may be coming unraveled. If so, it would be a funeral for the Great Moderation.

Is it really true that the Great Moderation is coming to an end? My sense is that it is too early to

James B. Bullard is president of the Federal Reserve Bank of St. Louis. The author appreciates the assistance and comments provided by colleagues at the Federal Reserve Bank of St. Louis. Marcela M. Williams, special research assistant to the president, provided assistance.

Figure 1

Real GDP Growth, 1950:Q1–2008:Q3



SOURCE: Bureau of Economic Analysis and National Bureau of Economic Research.

tell. Let's begin with a description of why policymakers and academics started talking about moderation and resiliency in the first place. The main idea is simple: Our primary measures of macroeconomic performance have been a lot less volatile than they were before 1984. In particular, quarterly gross domestic product (GDP) growth rates for the U.S. economy since 1945 show a clear change in behavior beginning in the middle 1980s. After 1984, these growth rates are only about half as volatile as they were during the earlier period.[1] So, for the past 25 years, growth rate volatility has dramatically moderated from what it was in the 1950s, 60s, and 70s (Figure 1).

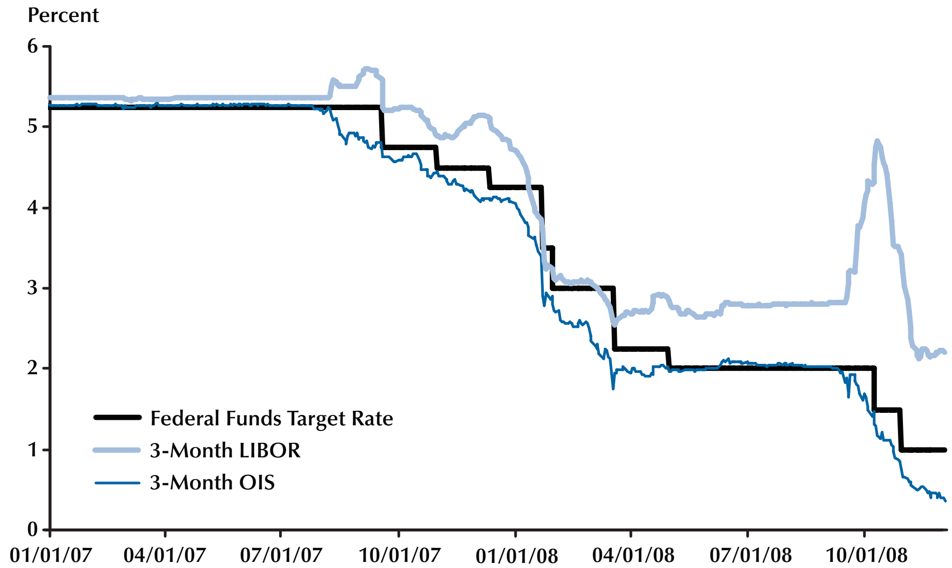
Furthermore, this phenomenon is not limited to real GDP growth rates. Almost all macroeconomic data have been dramatically less volatile since the mid-1980s, according to academic research (Stock and Watson, 2003). So the Great Moderation is a clear feature of the U.S. macroeconomic data since the mid-1980s. And, as is often the case when the data show a clear pattern, theories abound about the causes of this phenomenon. But all the theories have a common theme—namely, that some important macroeconomic event triggered a more stable, more resilient American economy over the past 25 years.

Understandably, many are yearning for a sense of stability today, and many are questioning what happened to the resiliency and moderation in the U.S. economy. Two areas stand out where volatility has been particularly high since the current financial turmoil began in earnest in August 2007. One is in certain interest rates and interest rate spreads, especially in markets that have experienced severe difficulties since the turmoil began. The closely watched LIBOR–Overnight Index Swap spread, for instance, peaked at more than 300 basis points before retreating in recent weeks. In July 2007, this spread was less than 10 basis points (Figure 2). Another volatile area is the equity markets: The Wilshire 5000 stock price index, one of the broadest measures of equity valuation, has been trading near its lows of 2002

1 For a discussion and some theorizing about the Great Moderation, see Bullard and Singh (2007).

Figure 2

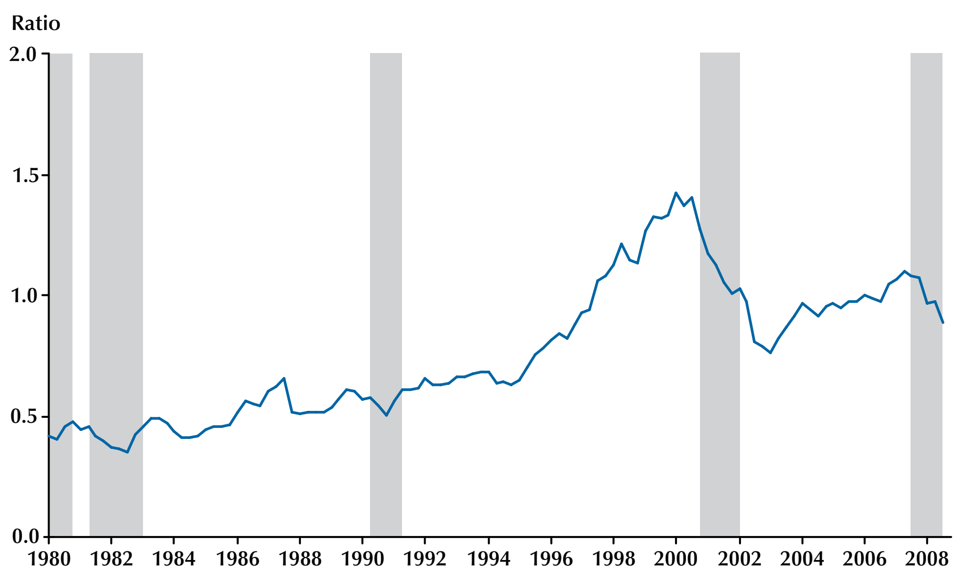
LIBOR, OIS, and Federal Funds Target Rate, January 3, 2007–December 10, 2008



SOURCE: Federal Reserve Board, British Bankers' Association, Reuters.

Figure 3

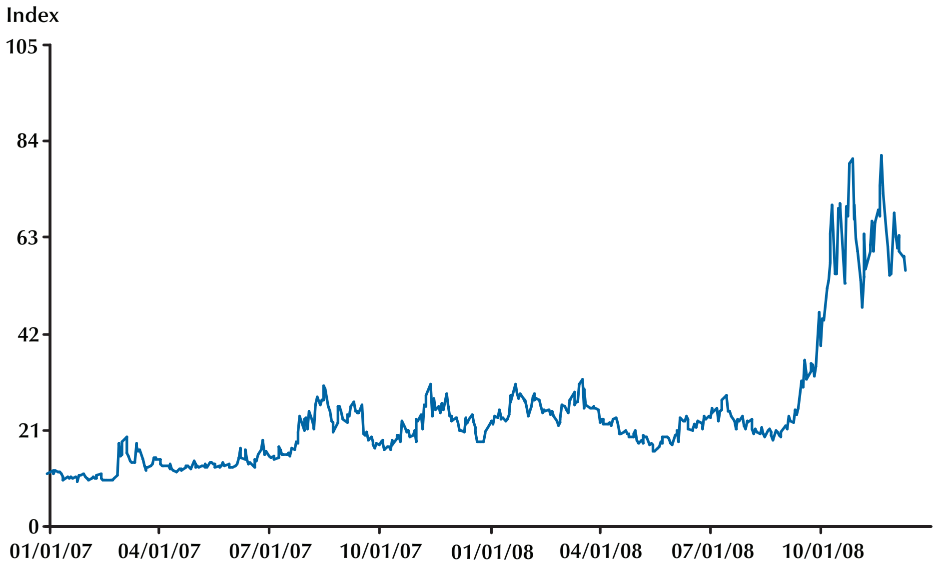
Wilshire 5000 Price Index/Nominal GDP, 1980:Q1–2008:Q3



SOURCE: *Wall Street Journal*, Bureau of Economic Analysis, and National Bureau of Economic Research.

Figure 4

Chicago Board Options Exchange VIX, January 3, 2007–December 10, 2008



SOURCE: *Wall Street Journal*.

and 2003 in the past few weeks (Figure 3). The Chicago Board Options Exchange market volatility index (VIX) was often above 60 during October and November of this year; in July 2007 it was below 20. The dramatic rise in volatility based on numbers like these is clear (Figure 4).

Still, it is far too early to organize a funeral for the Great Moderation. Even though financial market volatility is exceptionally high and the U.S. economy is contracting during the second half of 2008, the demise of the Great Moderation would require much more evidence than currently exists. Real economic variables, in particular, would have to swing much more than they have to date, and the increased volatility would have to continue for a number of years before we could start to compare the current environment with the pre-1984 experience and pronounce the moderation dead. Real GDP has fallen by half a percent at an annual rate in the third quarter of this year.[2] To be sure, fourth-quarter 2008 output is expected to fall sharply, followed by further but less-severe contraction in the first quarter of 2009. If that scenario materializes, the contour of the current recession will look much the same as that of the 1990-91 recession. As bad as that feels, it is not enough to undo 25 years of moderated behavior in the U.S. economy.

CHANGES IN THE FINANCIAL MARKETPLACE

It is no secret that the current financial market turmoil has brought about once-unimaginable changes on Wall Street. One telling sign of the magnitude of these changes is that the U.S. economy began 2008 with five large investment banks but will exit the year with zero. Without question, financial market turmoil since August 2007 is radically altering the nature of U.S. financial intermediation. I think it is fair to say that we are witnessing a funeral for the financial system we have known over the past two decades.

[2] The BEA preliminary estimate was released November 25, 2008.

A key culprit has been the illiquidity of mortgage-backed securities and related financial instruments. Many financial firms simply did not manage risk exposure on these securities well and as a result have struggled with losses and write-downs. The International Monetary Fund (2008) has estimated that more than $1.4 trillion of losses will have to be absorbed by the financial sector before all is said and done in this episode and that only a portion of these losses have been accounted for to date. The opacity of the financial instruments involved has kept everyone guessing as to where these losses truly lie, which explains a lot about how events have unfolded during 2008. No firm has an incentive to declare that it may suffer debilitating losses, and so markets have to discover which firms are insolvent and which are likely to survive and build market share in the post-shakeout industry structure. The sharp downturn in the real economy during the fall of 2008 has intensified the pressure. In the meantime, firms have become wary of trading with one another, certain markets have ceased normal functioning, and market participants and policymakers alike have been confronted with a series of announcements from firms near bankruptcy. In a November 17 *New York Times* editorial, Treasury Secretary Paulson named the litany of firms experiencing "failure, or the equivalent of failures": Bear Stearns, IndyMac, Lehman Brothers, Washington Mutual, Wachovia, Fannie Mae, Freddie Mac, and the American International Group (Paulson, 2008).

The Federal Reserve has been forced to improvise in response to firms' announcements of this nature. The key concern has been that, if important financial market players are failing, the failure should occur in an orderly way with the lowest level of market disruption. In the banking sector, there are well-established procedures for resolving a failed institution in an orderly way. These procedures have served the nation well both during the current crisis and during the savings and loan episode during the late 1980s and early 1990s. It is very important to recognize that there are no such procedures in the non-bank financial sector today. This regulatory gap is likely to be a primary focal point for discussions of the future of financial market regulation. In particular, any reform has to address the question of whether—and how—to set up systems to resolve failing non-bank financial firms in an orderly way. The current system—bankruptcy court—is not working.

As the shakeout process has unfolded during 2008, markets have been continually bracing for further surprise announcements from financial firms. The policy response to this situation has been exceptionally aggressive (Table 1). Consider the largest S&P 500 financial firms by assets as of the fourth quarter of 2007. The first 47 firms on the list accounted for 95 percent of the total assets held by the sector as of the fourth quarter of 2007. As of mid-summer 2008, just one of these financial firms had been the focus of a direct policy response of any kind. That firm was Bear Stearns, which, back in March, was purchased by JPMorgan Chase with help from the Fed. Almost all the others were operating as they had during recent years. The situation is dramatically different today. At the time of this writing, 22 of the 47 have received capital injections under the Treasury's Troubled Assets Relief Program (TARP) effort. Three of these are non-bank financial firms that changed their charters to become bank holding companies, including two of the largest firms on the list, Goldman Sachs and Morgan Stanley, and just recently American Express. Several other firms on the list merged with stronger partners, including Countrywide Financial and Merrill Lynch (both acquired by Bank of America), National City (acquired by PNC), Wachovia (acquired by Wells Fargo), and Sovereign Bancorp (approved acquisition by Banco Santander). The assets and debt obligations of Washington Mutual were purchased by JPMorgan Chase. Lehman Brothers went to bankruptcy court, but important portions of the company were acquired by Barclay's Capital. Fannie Mae and Freddie Mac were placed into conservatorship. American International Group has a restructured loan arrangement with the Treasury and the Federal Reserve. Citigroup has entered into an agreement to receive a package of guarantees, liquidity access, and capital from the government. These events have touched 33 of the 47 firms on the list. This means that much of the uncertainty surrounding the fate of U.S. finan-

Table 1

Status of Large S&P 500 Financial Firms

Firm	1-year percent change in stock price 11/30/07–11/28/08	Total assets ($ billions) (2007:Q4)	Percent of total assets in S&P 500 financials	Cumulative percent
Citigroup Inc.	–75%	$2,187.63	10.93%	10.93%
Bank of America Corp.	–65	1,715.75	8.57	19.50
JPMorgan Chase & Co.	–31	1,562.15	7.80	27.31
Goldman Sachs Group	–65	1,119.80	5.59	32.90
American International Group	–97	1,060.51	5.30	38.20
Morgan Stanley	–72	1,045.41	5.22	43.42
Merrill Lynch	–78	1,020.05	5.10	48.52
Fannie Mae	–97	882.55	4.41	53.93
Federal Home Loan Mtg.	–97	794.37	3.97	56.90
Wachovia Corp.	–87	782.90	3.91	60.81
Lehman Bros.	–100	691.06	3.45	64.26
Wells Fargo	–11	575.44	2.88	67.14
MetLife Inc.	–56	558.56	2.79	69.93
Prudential Financial	–77	485.81	2.43	72.35
Hartford Financial Services Group	–91	360.36	1.80	74.16
Washington Mutual	N/A	327.91	1.64	75.79
U.S. Bancorp	–18	237.62	1.19	76.98
Countrywide Financial Corp.	N/A	211.73	1.06	78.04
Bank of New York Mellon Corp.	–37	197.66	0.99	79.03
Lincoln National	–78	191.44	0.96	79.98
SunTrust Banks	–55	179.57	0.90	80.88
Allstate Corp.	–50	156.41	0.78	81.66
SLM Corporation	–76	155.56	0.78	82.44
Principal Financial Group	–79	154.52	0.77	83.21
Capital One Financial	–35	150.59	0.75	83.96
National City Corp.	–90	150.37	0.75	84.71
American Express	–60	149.83	0.75	85.46
State Street Corp.	–47	142.54	0.71	86.17
Regions Financial Corp.	–61	141.04	0.70	86.88
PNC Financial Services	–28	138.92	0.69	87.57
BB&T Corporation	–17	132.62	0.66	88.24
The Travelers Companies Inc.	–18	115.22	0.58	88.81
Genworth Financial Inc.	–94	114.32	0.57	89.38

SOURCE: Securities and Exchange Commission, Standard & Poor's, Federal Reserve Board, *Wall Street Journal*, and Government Accountability Office.

Type of firm	Current status (as of 12/11/2008)	Capital from government investment plan ($ billions) as of 12/11/08
BHC	According to 11/24/08 plan: Treasury and FDIC backstop of $300 billion in troubled assets; additional $20 billion stake in the firm by the Treasury.	45
BHC	Acquired Countrywide Financial and Merrill Lynch.	15
BHC	Acquired Bear Stearns. Acquired Washington Mutual's secured debt obligations and deposits.	25
BHC	Has become a BHC.	10
Insurance	Restructured plan, as of 11/10/08: $4 billion equity stake by government, $30 billion in funds on securities underlying the firm's CDS, $22.5 billion to buy residential mortgage securities. It will reduce the previous credit line to $60 billion.	40
BHC	Has become a BHC.	10
Inv. bank	Acquired by Bank of America.	10
GSE	Placed into conservatorship.	—
GSE	Placed into conservatorship.	—
BHC	Bought by Wells Fargo.	—
Inv. bank	Filed for bankruptcy; Barclay's has acquired important pieces.	—
Thrift	Bought Wachovia.	25
Insurance	—	—
Financial adv./insurance	—	—
Insurance	—	—
Thrift	Assets and debt obligations bought by JPMorgan Chase.	—
BHC	—	6.599
Thrift	Merged into Bank of America.	—
BHC	—	3
Insurance	—	—
BHC	—	3.5
Insurance	—	—
Credit services	—	—
Financial adv./asset mgmt.	—	—
BHC	Announced agreement to purchase Chevy Chase Bank.	—
BHC	Bought by PNC Financial.	3.555
BHC	Converted to BHC on 11/10/08, requested access to $3.5 billion TARP funds on 11/12/08.	—
BHC	—	2
BHC	—	3.5
BHC	Acquired National City Corporation on 10/24/08.	—
BHC	—	3.134
Insurance	—	—
Insurance	—	—

Table 1, cont'd

Status of Large S&P 500 Financial Firms

Firm	1-year percent change in stock price 11/30/07–11/28/08	Total assets ($ billions) (2007:Q4)	Percent of total assets in S&P 500 financials	Cumulative percent
Fifth Third Bancorp	–68%	$110.96	0.55%	89.94%
Ameriprise Financial Inc.	–69	109.23	0.55	90.48
KeyCorp	–64	99.98	0.50	90.98
Bear Stearns Cos.	N/A	96.08	0.48	91.46
CIT Group	–87	90.25	0.45	91.91
Sovereign Bancorp	–79	84.75	0.42	92.34
Loews Corp.	–43	76.08	0.38	92.72
ACE Limited	–13	72.09	0.36	93.08
Northern Trust Corp.	–43	67.61	0.34	93.41
AFLAC Inc.	–26	65.81	0.33	93.74
M&T Bank Corp.	–29	64.88	0.32	94.07
E*Trade Financial Corp.	–71	64.19	0.32	94.39
Comerica Inc.	–51	62.33	0.31	94.70
Marshall & Ilsley Corp.	–51	58.30	0.29	94.99

SOURCE: Securities and Exchange Commission, Standard & Poor's, Federal Reserve Board, *Wall Street Journal*, and Government Accountability Office.

cial companies has been addressed in one manner or another during the past several months. Turmoil is still significant, to be sure, but the policy response has been very aggressive, and many of the largest uncertainties have been addressed.

FUNERAL FOR A FRIEND

The financial market turmoil began during the summer of 2007. The initial analysis—widely accepted among policymakers and forecasters—suggested that problems in the subprime sector of the mortgage markets were unlikely to be large enough to have a substantial impact on the U.S. economy outside the financial sector. Markets seemed to confirm this judgment during the fall of 2007, as equity indexes peaked. By late 2007, however, it became apparent that problems were going to be more difficult and long-lasting than the original analysis suggested. Real GDP growth in the fourth quarter of 2007 was slightly negative based on the revised data available today. Employment growth turned negative in January 2008.

The Fed responded to the weakening economy by easing aggressively, lowering the target federal funds rate by 225 basis points during the first few months of 2008 all the way down to 2 percent. During the summer, the Federal Open Market Committee (FOMC) went on hold, but intensified turmoil during the fall combined with weaker-than-expected data on the real economy triggered further easing moves during October. This has left the FOMC with a federal funds target at a low level, with further easing possible as weak data roll in over the next several months (Figure 5A).

Whether the FOMC decides to stay on hold at this point or eases further and then stays on hold at some lower level, even zero, may not be the most critical question. The fact is, monetary policy defined as movements in short-term nominal interest rates is coming to an end, at least for now. It's a funeral for a friend.

The end of nominal interest rate targeting in the United States for the near term means that

Type of firm	Current status (as of 12/11/2008)	Capital from government investment plan ($ billions) as of 12/11/08
BHC	—	3.45
Financial adv.	—	—
BHC	—	2.5
Inv. bank	Fed arranged merger with JP Morgan Chase.	—
Credit services	—	—
Thrift	Approved takeover by Banco Santander 10/13/08	—
Insurance/hotels	—	—
Insurance	—	—
BHC	—	1.576
Insurance	—	—
BHC	—	0.6
Inv. brokerage	—	0.8
BHC	—	2.25
BHC	—	1.715

much more attention will have to be paid to alternative ideas about controlling inflation and inflation expectations going forward. An important characteristic of the current environment is that medium-term inflation expectations seem to be spreading out dramatically, with some analysis warning of high inflation, others warning of deflation, and still others expecting inflation to remain near the levels recently experienced.

One focus of analysis over the coming quarters will be the experience in Japan. Japan was buffeted by large declines in equity and real estate markets in the early 1990s. In response, the Bank of Japan lowered nominal interest rates to near zero by the middle of that decade, and it has not been above 1 percent since. An important part of the outcome in Japan has been a rate of deflation that has averaged about 1 percent since the mid-1990s (Figure 5B). Deflation, should it occur in the United States, might be particularly challenging because some of our current core problems are in housing markets, where contracts are written in nominal terms. An unexpected deflation would make those contracts more expensive for borrowers.

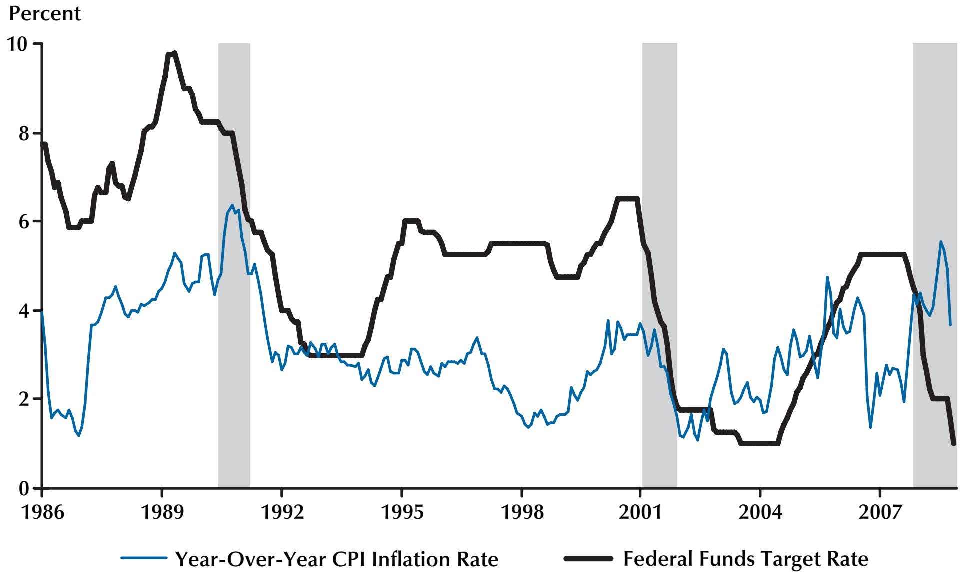
One idea from the Japanese experience is that with nominal interest rates at very low levels, more attention may have to be paid to quantitative measures of monetary policy. By announcing and maintaining targets for key monetary quantities, the Fed may be able to keep inflation and inflation expectations near target and ward off either a drift toward deflation or excessively high inflation. This will be an important issue for the Fed in coming months and represents a challenge in the communication of monetary policy going forward.

A REBIRTH

So far, I have discussed three funerals, ideas whose times may have passed. I now want to turn to a macroeconomic idea that is being rehabilitated as we speak. That idea is fiscal policy—in particular, the spending side of fiscal policy and the idea of more direct intervention in the affairs of private sector firms.

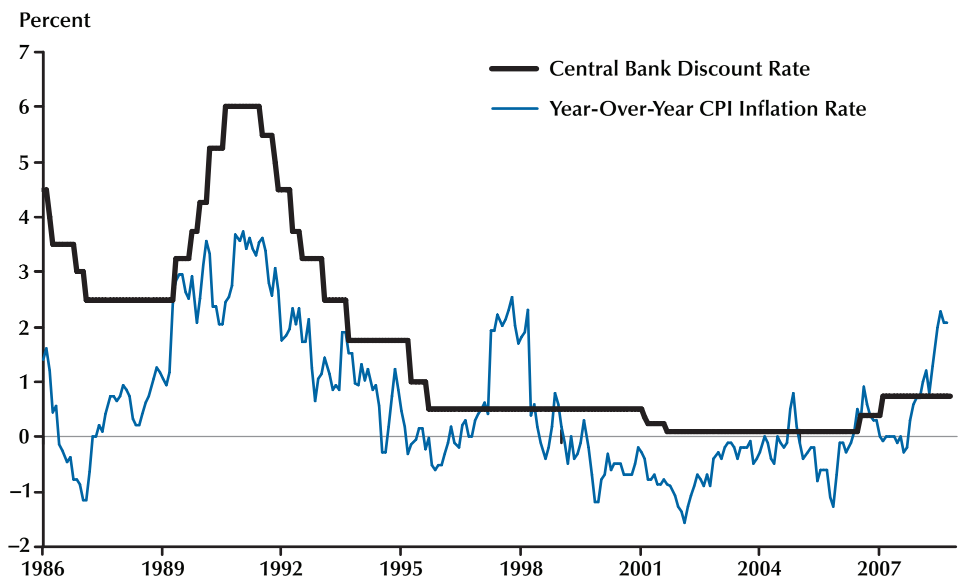
Figure 5

A. Federal Funds Rate and CPI Inflation Rate, January 1986–November 2008



SOURCE: Federal Reserve Board, Bureau of Labor Statistics, National Bureau of Economic Research.

B. Japan Policy Rate and Year-Over-Year Inflation Rate, January 1986–October 2008



SOURCE: Organisation for Economic Co-operation and Development.

At least since the 1980s, fiscal policy defined by deficit spending has had a negative connotation in many macroeconomic policy circles. Former Council of Economic Advisers Chairman N. Greg Mankiw, writing in 1991, listed as his "dubious Keynesian proposition #4" the idea that "fiscal policy is a powerful tool for economic stabilization." He included this sentence: "In the United States today, fiscal policymakers have completely abdicated responsibility for economic stabilization." That was 1991, but I think it is a fair assessment of the thinking in much of the economics profession up until the current financial market turmoil. Fiscal policy, at least in the United States, was viewed as important for the macroeconomy, but from a longer-run perspective. To the extent there are stabilization goals—goals requiring time-critical policy interventions—the usual idea is that certain types of tax cuts might be beneficial, but that otherwise the effort is best left to monetary policy. Not least in this thinking is that the Fed can act relatively quickly, while the political process tends to be much slower and more cumbersome.

Yet, during the fall of 2008 in particular, fiscal policy conceived of as more direct intervention in the operation of private sector firms has emerged as a leading tool to combat ongoing financial market turmoil. This is so, not just in the United States, but globally.[3] The passage of the Emergency Economic Stabilization Act (EESA), with authorization for the Treasury to spend up to $700 billion to help return financial markets to more normal operation, has put the focus going forward squarely on fiscal responses. This created in a matter of weeks a very different policy environment from the one that had existed in the United States for the past 25 years.

The original idea behind the EESA was to create a market for the illiquid asset-backed securities and related instruments that are at the heart of the present situation. These assets have current prices, to the extent that they can be determined, that are very low, the so-called fire sale price, because so many firms would like to sell their holdings and few buyers exist in the current climate. However, these securities also have a higher, hold-to-maturity price that reflects the likely value of the stream of revenue for a patient investor who is willing to simply hold the asset for a period of time. Under the original EESA proposal, the government would play the role of the patient investor, buying the securities at a reverse auction and holding them or selling them at a future moment when financial market stress has receded. In principle, this idea could be executed at no ultimate cost to the taxpayer, although taxpayer money would be put at risk.[4] An important part of the concept is that taxpayer money would be used to purchase assets that would then be sold in the future, recouping most or all of the initial outlay. The government would not have to purchase all assets, only enough to credibly create a market. I thought such a program, if it could be executed on a sufficient scale, may have helped to liquefy illiquid asset-backed securities markets and so may have helped progress toward an orderly financial market consolidation. This, in turn, would have helped to reduce or eliminate the downside risk to economic performance.

As events have transpired, the asset-purchase program has been put on hold. Given the rapid flow of events, capital purchases came to be viewed during the autumn as a simpler, more timely, and more direct method of intervention. The Treasury's capital injection program has taken the bulk of the resources from the first $350 billion tranche of the $700 billion appropriation.

CONCLUSIONS

I have described three funerals and a wedding. The ongoing financial market turmoil may have caused the death of many cherished ideas about

[3] International Monetary Fund (2008, "Recent Central Bank and Government Actions," pp. 7-10).

[4] One place to look for a model for handling financial crises of this magnitude is the Nordic countries during the early 1990s. For a recent summary, see the speech by my friend and colleague Seppo Honkapohja, a governor at the Bank of Finland (Honkapohja, 2008). These countries were hit by severe financial turmoil and sharp recessions, in part associated with currency crises, in the early 1990s. The general response was for the governments to take equity positions in banks and to manage the resulting consolidation in the industry. As Honkapohja documents, the ultimate expense to the taxpayers in these countries was less than the initial outlay of government funds.

how the macroeconomy operates. One funeral was for the idea of the Great Moderation. Certainly financial markets have seen exceptional volatility recently, and some behavior in those markets has been unprecedented. Still, I am not ready to bury the Great Moderation yet—we will need a lot more very volatile data on the real side of the economy to truly depart from the experience of the past 25 years. A second funeral was for our financial system as we have known it. That transformation has occurred and continues, with repercussions for U.S. and global financial market regulation. A third funeral was for monetary policy defined as nominal interest rate targeting. At least over the near term, any additional influence through interest rate reductions will be limited, and the focus of monetary policy may turn to quantity measures. The wedding—the idea on the rise—is fiscal policy defined as more direct intervention in certain parts of the private sector. While the Fed will continue to be innovative in providing liquidity to markets through existing facilities and possibly some new programs, an important part of the response to ongoing financial market turmoil will come from fiscal policy intervention. This runs counter to much of the thinking in macroeconomic policy circles over the past two decades. It may be discomforting or rewarding or both, but stabilization policy in the coming months and quarters is likely to look very different from what we have been accustomed to seeing.

REFERENCES

Bullard, James B. and Singh, Aarti. "Learning and the Great Moderation." Working Paper 2007-027a, Federal Reserve Bank of St. Louis, June 2007; research.stlouisfed.org/wp/2007/2007-027.pdf.

Hoenig, Thomas M. "Global Financial Crisis and the Regulatory Environment: Where Do We Go From Here?" Presented at the Institute of International Bankers, New York, November 17, 2008; www.kansascityfed.org/SpeechBio/HoenigPDF/Hoenig.Global.11.17.08.pdf.

Honkapohja, Seppo. "The 1990's Financial Crises in Nordic Countries." Presented at the Global Interdependence Center, Philadelphia, September 28, 2008; www.bof.fi/NR/rdonlyres/055DF411-915C-4A0F-8B1C-B756928E319F/0/ 080928_SH_Nordic_model.pdf or www.interdependence.org/docs/The_1990_Nordic_Crisis_Seppo_Honkapohja.pdf.

International Monetary Fund. "Global Financial Stability Report: Financial Stress and Deleveraging—Macrofinancial Implications and Policy," Washington, DC: IMF, 2008; www.imf.org/external/pubs/ft/gfsr/2008/02/pdf/text.pdf.

Mankiw, N. Gregory. "The Reincarnation of Keynesian Economics." NBER Working Paper No. W3885, National Bureau of Economic Research, October 1991 (issued July 1992); www.nber.org/papers/w3885.pdf.

Paulson, Henry M. Jr. "Fighting the Financial Crisis, One Challenge at a Time." *New York Times*, November 17, 2008; www.nytimes.com/2008/11/18/opinion/18paulson.html.

Plosser, Charles I. "Some Thoughts on the Economy and Financial Regulatory Reform." Presented at The Economics Club of Pittsburgh, November 13, 2008; www.philadelphiafed.org/publications/speeches/plosser/2008/11-13-08_economics-club-of-pittsburgh.pdf.

Stock, James H. and Watson, Mark W. "Has the Business Cycle Changed and Why?" in Mark Gertler and Kenneth S. Rogoff, eds., *NBER Macroeconomics Annual 2002*. Volume 17. Cambridge, MA: MIT Press, 2003, pp. 159-218.

The Fed, Liquidity, and Credit Allocation

Daniel L. Thornton

The current financial turmoil has generated considerable discussion of liquidity. Moreover, it has been widely reported that the Federal Reserve played a major role in supplying liquidity to financial markets during this distressed time. This article describes two ways in which the Fed has supplied liquidity since late 2007. The first is traditional: The Fed supplies liquidity by providing credit through open market operations and by lending to depository institutions at the so-called discount window. The second is by enhancing the liquidity of portfolios of some institutions by replacing their less-liquid assets with more-liquid assets. The Fed has used the second approach since late 2007. Unlike several previous occasions, however, it began supplying liquidity in the first, more traditional way only recently—in September 2008. This article notes that the Fed departed from its long-standing tradition of minimizing its effect on the allocation of credit by supplying liquidity to institutions that it believed to be most in need; at the same time, it neutralized the effects of these actions on the total supply of liquidity in the financial market. The article also discusses the Fed's reasons for reallocating credit this time rather than simply increasing the total supply of financial market liquidity. (JEL E44, E52, E58)

Federal Reserve Bank of St. Louis *Review*, January/February 2009, *91*(1), pp. 13-21.

ASSET LIQUIDITY AND FINANCIAL MARKET LIQUIDITY

Unfortunately, the word "liquidity" is often used to describe very different things. Liquidity is perhaps most often used to describe a particular characteristic of an asset. In this sense, liquidity means the "degree of ease and certainty of value with which a security can be converted into cash." Cash is pure liquidity. Every other asset has a degree of liquidity that is determined by (i) how quickly it can be converted to cash and (ii) how much the price of the asset must be reduced to do so. The second requirement stems from the fact that virtually any asset can be converted to cash quickly if the price is sufficiently attractive.

The word "liquidity" is also used to describe the availability of credit in the financial market. For example, market analysts or policymakers might say there is a shortage of liquidity in the market or that the financial market is "frozen up." This means that it is difficult or expensive to obtain a loan (i.e., get credit). Like the liquidity of an asset, this concept of market liquidity is relative. Even in the most liquid of financial markets, some individuals or firms will be unable to obtain a loan or, if they do, they will be charged a relatively high interest rate. Likewise, many individuals or institutions obtain credit in markets described as "illiquid." No absolute measure of the liquidity of the financial market exists.

An important distinction separates the concept of market liquidity from the concept of asset

Daniel L. Thornton is a vice president and economic adviser at the Federal Reserve Bank of St. Louis. The author thanks Aditya Gummadavelli and Mary Karr for research assistance.

liquidity. By the latter definition, cash is the quintessence of liquidity; however, "a shortage of liquidity" in the financial market does not mean a shortage of cash because there can never be a shortage of cash. This was not always the case. Before the establishment of the Federal Reserve, shortages of cash did occur. However, when the Federal Reserve was established, it was designed to provide an "elastic currency." That is, it was designed so that the quantity of cash automatically increases to meet society's demand for it: Thus, there can never be a shortage of cash. When market analysts and others say that the market has become less liquid or is illiquid, they mean that it is more difficult to get a loan than before; they do not mean there is a shortage of cash.

THE FED AS A SUPPLIER OF MARKET LIQUIDITY

Fundamentally, domestic credit has three major sources: private saving (individuals and firms), government saving (surpluses of federal, state, and local governments), and changes in the monetary base—the sum of cash held by the public and bank reserves. The Fed supplies the market with credit through open market operations and, to a much lesser extent historically, through loans to depository institutions at the discount window. These actions increase the total supply of credit in the financial market. This is most easily seen in Fed lending at the discount window. When the Fed makes a loan at the discount window, it is directly extending credit to the borrowing institution. That is, the Fed takes the IOU of the borrowing institution in return for funds—specifically, deposit balances at the Fed.

The effect of an open market purchase of securities on the total supply of credit is exactly the same as an equal amount of lending at the discount window. In this case, the Fed acquires a security (i.e., an IOU) in exchange for funds—deposit balances at the Fed. Historically, the Fed has conducted open market operations in government and agency securities; however, open market operations can be carried out in any asset prescribed by the Federal Reserve Act. When the Fed purchases Treasury securities from the public, it is indirectly making the loan to the Treasury rather than the public. Hence, the supply of credit available to the public increases.[1] Of course, if the Fed sells some of its securities, the supply of credit available to the public declines. All other things equal, the supply of credit in the financial market increases or decreases as the monetary base increases or decreases, regardless of whether the change in the monetary base is due to Fed lending or open market operations.

THE FED AND THE ALLOCATION OF CREDIT

Although lending by the Fed has exactly the same effect on the monetary base as an equivalent open market operation, the effect of these actions on the allocation of credit is different. When the Fed makes a loan to a depository institution, or anyone else, it directly allocates credit to that institution. The effect on the allocation of credit is mitigated by the fact that the total supply of credit increases—the borrowing institution obtains credit and no one loses credit. The effect of Fed lending on the allocation of credit is intensified when the Fed offsets the effect of its lending activity on the total supply of credit through open market operations. In this case, the borrowing institution obtains credit but the total supply of credit is unchanged. In effect, the borrowing institution is getting credit at the expense of some other individual or institution: The total supply of credit is reallocated.

Historically, the Fed has offset the effect of discount window lending on the total supply of credit through open market operations.[2] That is, if depository institutions borrowed at the discount window, the Fed would offset the effect of this increased borrowing on the monetary base by selling a comparable amount of securities in an open market operation.

[1] This example is used for ease of understanding. The effect is the same regardless of what the Fed purchases.

[2] For a dramatic example that had important implications for monetary policy analyses, see Thornton (2001).

The practice of offsetting the effect of discount window lending on the monetary base means that discount window lending reallocated credit to the borrowing institution. The effect of discount window lending on credit allocation has not been an issue for two reasons. First, the initial effect of an open market operation is on depository institutions. Consequently, a discount window loan to a depository institution that is offset through open market operations has the effect of reallocating credit among depository institutions.

Second, and more important, discount window lending has been small historically. For much of its history, the Fed has discouraged depository institutions from borrowing at the discount window. Depository institutions were expected to come to the window only when they had exhausted the relevant alternative sources of funds. In addition, following the substantial borrowing by then-troubled Continental Illinois Bank in May 1984, depository institutions grew increasingly reluctant to borrow from the Fed because of concern that such borrowing institutions would be perceived as "troubled."[3]

For these reasons discount window borrowing has been small historically. For example, from January 1985 though December 2007, discount window borrowing averaged $547 million—less than two-tenths of 1 percent of the monetary base. Consequently, borrowing has had little effect on the allocation of credit in the financial market. Moreover, because of the Fed's practice of offsetting the effect of borrowing, discount window borrowing has had little effect on the monetary base. The correlation between discount window borrowing and changes in the monetary base from January 1985 through August 2008 was essentially zero (less than 1 percent).

The insignificant effect of such borrowing on the allocation of credit in the financial market is consistent with the Fed's long-standing practice of minimal interference in the government securities market in particular and the credit market more generally. The Fed traditionally has conducted open market operations at the very short end of the maturity structure and primarily in Treasury securities to minimize the effect of its operations on the structure of interest rates. With the exception of a short departure in the early 1960s, this policy has guided the conduct of open market operations.[4]

THE FED'S NEW LENDING FACILITIES AND THE ALLOCATION OF CREDIT

In response to the distress in financial markets associated with the decline in house prices, the Fed initiated a series of new lending programs that according to Cecchetti (2008) were implemented to ensure "that liquidity would be distributed to those institutions that needed it most."[5] First among these programs was the Term Auction Facility (TAF), by which the Fed auctions funds to depository institutions. The TAF differs from normal discount window borrowing in two respects. First, rather than coming to the Fed to request a discount window loan, under the TAF the Fed auctions a predetermined amount of funds. Second, rather than paying the "primary credit rate" (formerly known as the discount rate), depository institutions that borrow under the TAF pay the "stop-out rate"—the lowest bid rate that exhausts the funds being auctioned.[6] It was hoped that the TAF's alternative method of borrowing would counter depository institutions' reluctance to borrow from the Fed. Once depository institutions became comfortable with borrowing from the Fed, the stigma associated with discount window borrowing would be reduced. This appears to have happened. Primary credit borrowing averaged $11.85 billion during the first nine months of 2008. However, Thornton (2008) shows that depository institutions borrow at the discount

[3] See Thornton (2001) for a discussion of the effects from the Continental Illinois Bank experience on discount window borrowing.

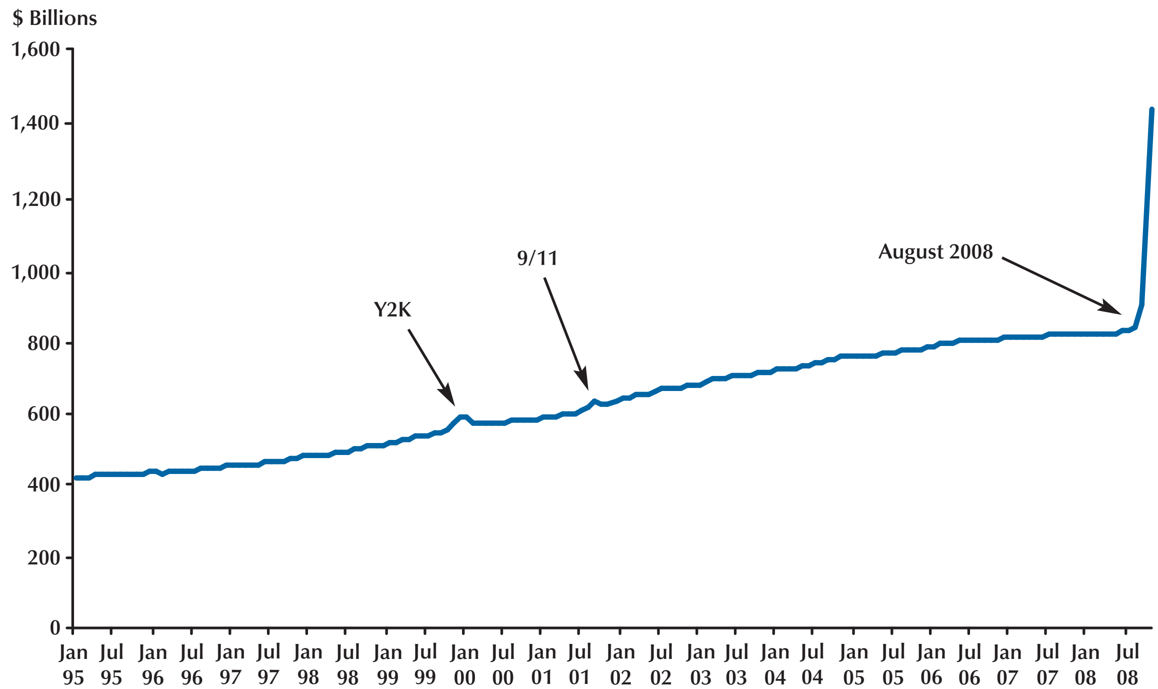
[4] For discussions of this so-called bills-only policy, see Friedman and Schwartz (1963) and Meltzer (2009). The classic article on the Fed's brief deviation from this policy, called "Operation Twist," is by Modigliani and Sutch (1966).

[5] Cecchetti (2008, abstract).

[6] The Fed establishes a minimum bid rate at which it will lend. Loans are made at the minimum bid rate only when the demand for loans at this rate is less than or equal to the amount being auctioned.

Figure 1

Monetary Base (January 1995–November 2008)



window only when the primary credit rate is lower than the rates on alternative sources of funds.

The Fed subsequently initiated several additional lending facilities. The Primary Dealer Credit Facility (PDCF) essentially opened the discount window to primary government security dealers.[7] The Asset-Backed Commercial Paper Money Market Mutual Fund Liquidity Facility (ABCP MMMF Liquidity Facility) is intended to increase liquidity in the commercial paper market by providing loans to U.S. depository institutions and bank holding companies for the purpose of purchasing high-quality asset-backed commercial paper.[8]

Under the TAF, the PDCF and, most recently, the ABCP MMMF Liquidity Facility, the Fed is essentially making loans to the participating institutions. All other things equal, such loans increase the monetary base. Until mid-September 2008, the Fed offset the effect of these lending programs on the total supply of credit through open market operations. Figure 1 shows the level of the monetary base from January 1995 through November 2008. The figure shows that the hundreds of billions of dollars of "liquidity" supplied through these facilities had no impact on the monetary base and, consequently, no effect on the total supply of credit in the financial market until September 2008.

The Fed's behavior of not increasing the total supply of credit when there were liquidity concerns differs markedly from its response to liquid-

---

[7] For a list of these dealers, access www.newyorkfed.org/markets/pridealers_current.html.

[8] There are several other lending facilities not discussed here. For more information on these new lending facilities established before May 2008, see Cecchetti (2008). For information on all of the new lending facilities, visit the websites of the Federal Reserve Bank of New York or the Board of Governors of the Federal Reserve System. These new lending facilities are temporary; however, there has been some discussion about making the TAF permanent.

ity concerns on two previous occasions. Figure 1 shows two prior occasions when the monetary base increased sharply. The first occurred in late 1999 and was associated with Y2K—that is, widespread concerns about computer failures associated with the century date change. Such worries included beliefs that Y2K changes might significantly reduce the liquidity of the financial market. To guard against this possibility, the Fed injected relatively large amounts of base money (i.e., credit) through open market operations. The Y2K concerns never materialized. With no need for additional liquidity, the Fed quickly drained the base money it had supplied in anticipation of a liquidity shortage, and the monetary base resumed its normal growth path.

The second instance of liquidity influx by the Fed was associated with the terrorist attacks on September 11, 2001. Financial institutions that occupied the World Trade Center played an important role in U.S. financial markets. The terrorist attack on the World Trade Center significantly impeded the operations of these institutions and, importantly, their ability to provide credit. Recognizing this liquidity shortage, the Fed responded quickly and increased the monetary base by well over $100 billion. The affected firms were able to resume more or less normal operations quickly, so the additional base money was only supplied for a few days.[9]

Despite claims that the Fed has been supplying massive amounts of liquidity through its new lending programs, Figure 1 shows that no sharp rise in the monetary base occurred until September 2008; the liquidity supplied by the Fed was being offset through open market operations. Hence, the Fed did not increase the total supply of liquidity to the financial markets, as it did for Y2K or 9/11. These facilities merely increased the liquidity of the participating institutions' balance sheets by allowing participating institutions to exchange less-liquid (or illiquid) assets for highly liquid assets. This is particularly true of the Term Securities Lending Facility (TSLF) through which primary dealers essentially borrow specific Treasury securities offered by the Fed in exchange for less-liquid securities. These loans have no potential to increase the monetary base because they are essentially an exchange of less-liquid assets of the government security dealers for more-liquid Treasury securities held by the Fed.

The Fed's action to offset the effect of this borrowing on the supply of liquidity suggests that these facilities were intended only to increase the liquidity of the participating institutions' balance sheets, without increasing the liquidity of the financial market generally. In so doing, these programs had a significant effect on the allocation of credit by the Fed. As of November 19, 2008, the total amount of loans outstanding under the TAF, the other lending programs, and regular discount window borrowing was $1,611.5 billion, whereas the total monetary base was $1,476.4 billion. Hence, nearly all of the total credit supplied by the Fed was being allocated directly to participating institutions.

Beginning in September 2008, the Fed increased its total supply of credit to the market. Between August 27, 2008, and November 19, 2008, the monetary base increased by about $635.2 billion. Over this same period, Fed lending increased by $1,308.9 billion. Hence, the Fed offset the effect on the total supply of credit of 48.5 percent of its additional lending during the period.

CONVENTIONAL VERSUS UNCONVENTIONAL MONETARY POLICY

The Fed's response to liquidity concerns is a clear departure not only from its actions during Y2K and 9/11, but also from reliance on conventional tools of monetary policy. This current episode raises two interesting questions. Why did the Fed address the liquidity problem by creating a new array of lending programs rather than relying on conventional open market operations and the discount window? And why did the Fed decide to reallocate the total supply of credit rather than increase the total supply of liquidity in the financial market as it did for Y2K and 9/11?

[9] See Neely (2004).

Cecchetti (2008) suggests that the Fed instituted the new lending programs because it was not confident that it could allocate the credit to the financial institutions most in need of liquidity by using traditional tools. Specifically, he notes that "While well-established mechanisms existed for injecting reserves into a country's financial system, officials had no way to guarantee that the reserves will reach the banks that need them."[10]

Benanke (2008) appears to confirm Cecchetti's (2008) suggestion. While noting that the European Central Bank (ECB) and the Bank of England responded using conventional tools of monetary policy, Bernanke (2008) observed the following:

> In the United States, in ordinary circumstances only depository institutions have direct access to the discount window, and open market operations are conducted with just a small set of primary dealers against a narrow range of highly liquid collateral. In contrast, in jurisdictions with universal banking, the distinction between depository institutions and other types of financial institutions is much less relevant in defining access to central bank liquidity than is the case in the United States. Moreover, some central banks (such as the ECB) have greater flexibility than the Federal Reserve in the types of collateral they can accept in open market operations. As a result, some foreign central banks have been able to address the recent liquidity pressures within their existing frameworks without resorting to extraordinary measures. In contrast, the Federal Reserve has had to use methods it does not usually employ to address liquidity pressures across a number of markets and institutions. In effect, the Federal Reserve has had to innovate in large part to achieve what other central banks have been able to effect through existing tools.

Bernanke (2008) continues by suggesting that the

> traditional framework for liquidity provision was not up to addressing the recent strains in short-term funding markets. In particular, the efficacy of the discount window has been limited by the reluctance of depository institutions to use the window as a source of funding. The "stigma" associated with the discount window, which if anything intensifies during periods of crisis, arises primarily from banks' concerns that market participants will draw adverse inferences about their financial condition if their borrowing from the Federal Reserve were to become known.

Bernanke's (2008) statement suggests that the Fed was unable to direct the liquidity to institutions most in need using open market operations. However, the Federal Reserve Act (hereafter, Act) does not prevent the Fed from purchasing asset-backed securities, commercial paper, and a wide range of other securities, such as those taken as collateral against loans under the new lending programs.[11] Nor does the Act prevent the Fed from engaging in open market operations with institutions other than primary security dealers. Although the Fed would have had to modify its open market operating procedures, nothing in the Act per se would have prevented the Fed from using open market operations rather than an array of new lending programs to channel liquidity to institutions or markets most in need of liquidity.

Why the Fed chose not to increase the supply of total liquidity before September 2008 remains unclear. One possibility is that the Fed was concerned that massive injections of liquidity in the financial market would impair its ability to control the federal funds rate. Although he did not specifically state this as the reason, Bernanke (2008) noted that "open market operations have long been the principal tool used by the Federal Reserve to manage the aggregate level of reserves in the banking system and thereby control the federal funds rate."

---

[10] Cecchetti (2008, p. 15).

[11] Section 12A of the Act (www.federalreserve.gov/aboutthefed/section12.htm) created the Federal Open Market Committee (FOMC) and limited the authority of Federal Reserve Banks to undertake open market operations without FOMC direction. Section 14 of the Act (www.federalreserve.gov/aboutthefed/section14.htm) specifies the kinds of "normal course" paper that are used for open market operations. The list is exhaustive (see Small and Clouse, 2005). Open market operations are governed by FOMC rules outlined in 12 CFR 270 (http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&tpl=/ecfrbrowse/Title17/17cfr270_main_02.tpl), which limit the types of securities that the Fed can buy or sell in the normal course of operations. However, Section 270.4(d) of these regulations states that the "Federal Reserve Banks are authorized and directed to engage in such other operations as the Committee may from time to time determine to be reasonably necessary to the effective conduct of open market operations and the effectuation of open market policies."

Another reason may have prompted the Fed's unconventional approach. Benanke (2008) notes that "recent research by Allen and Gale (2007) confirms that, in principle at least, 'fire sales' forced by sharp increases in investors' liquidity preference can drive asset prices below their fundamental value, at a significant cost to the financial system and the economy." Bernanke goes on to say that "A central bank may be able to eliminate, or at least attenuate, adverse outcomes by making cash loans secured by borrowers' illiquid but sound assets."[12] Benanke (2008) suggests that in so doing borrowers could avoid selling securities in an illiquid market, which would avoid potential economic damage "arising, for example, from the unavailability of credit for productive purposes or the inefficient liquidation of long-term investments."

THE EFFICACY OF THE NEW APPROACH

Beyond the question of *why* the Fed chose this unconventional approach to monetary policy is the question of *how effective* it is. Many macroeconomists believe that changes in the composition of the Fed's assets that are not accompanied by a change in the monetary base are ineffective. This belief is due, in part, to experience. In the early 1960s, the Fed attempted to reduce long-term interest rates while maintaining relatively high short-term interest rates using a procedure called "Operation Twist." Specifically, the Fed bought long-term securities while simultaneously selling short-term securities, so that the net effect of these transactions on the monetary base was nil. The rationale was that by increasing the demand for long-term securities and reducing the demand for short-term securities, the Fed could "twist" the yield curve—long-term rates would fall relative to short-term rates. Most analysts concluded that the Fed had little or no effect on the shape of the yield curve.

Operation Twist's failure is consistent with alternative theories of the term structure. For example, the expectations hypothesis asserts that long-term rates are determined by the market's expectation of the future short-term rate. If short-term rates are not expected to fall, then long-term rates will not fall either. The failure of Operation Twist is also consistent with the risk-premium hypothesis, which suggests that rates on long-term securities are generally higher than rates on short-term securities because investors demand a risk premium for investing in longer-term securities because they have a higher degree of market risk. The risk premium is determined by what economists refer to as "deep structural parameters"—that is, the risk aversion of investors. A change in the relative demands for long-term and short-term securities has no effect on the size of the risk premium and, hence, no effect on the shape of the yield curve.

Similar experiences and theoretical arguments apply to attempts to alter the exchange rate through sterilized foreign exchange intervention. Sterilized foreign exchange intervention occurs when a central bank purchases securities denominated in one country's currency and simultaneously sells an equal amount of securities denominated in another country's currency, so the effect on the monetary base is nil. Theory and evidence suggest that foreign sterilized exchange intervention has little or no effect on exchange rates.

Considerable research will be done in the years to come to determine the efficacy of the Fed's new lending programs. Some early work by Taylor and Williams (2008a,b) indicates that the TAF was ineffective in significantly influencing the spread between term LIBOR rates (and other similar rates) and overnight lending rates, which started to rise dramatically in August 2007. Taylor and Williams suggest that the TAF was initiated in part to reduce the spread between term LIBOR rates and overnight lending rates. This motivation is supported by the Fed's February 2008 report to Congress, which states that, "although isolating the impact of the TAF on financial markets is not easy, a decline in spreads in term funding markets since early December provides some evidence that the TAF may have had beneficial effects on

[12] Bernanke (2008).

financial markets" (Board of Governors of the Federal Reserve, 2008).

Taylor and Williams (2008a,b) argue that the rate spread had increased as a result of banks' and other creditors' heightened reluctance to lend to banks perceived to have an increased risk of default. Hence, the rise in term LIBOR rates and other rates that reflect the cost of funds to banks, relative to overnight lending rates, reflects a risk premium that will not be reduced by increasing the liquidity of these banks' portfolios. Taylor and Williams (2008a) conclude that because "the TAF does not affect total liquidity, expectations of future overnight rates, or counterparty risk," it did not affect the rate spread.

CONCLUSION

In response to the financial turmoil in the wake of declining house prices, the Fed instituted a series of new lending facilities that increased the liquidity of participating institutions' portfolios without simultaneously increasing the total supply of liquidity in the financial market, at least before September 2008. In so doing, the Fed departed significantly from its historical practice of relying on traditional tools of open market operations and discount window lending to provide liquidity to the financial market.

Why the Fed chose to enact a series of new lending programs rather than use its existing tools of open market operations and the discount window is unclear. Given the stigma attached with borrowing from the discount window, the Fed would have had difficulty increasing the supply of total credit by making discount window loans. It could have increased the total supply of credit in the market through open market operations. However, the Federal Open Market Committee (FOMC) would have had to change its operating rules to purchase a broad array of securities, such as those it has taken as collateral under it new lending programs, and to engage in open market operations with entities other than primary security dealers.

It appears, however, that at least initially, the Fed did not want to address the financial market turmoil by increasing the total amount of credit in the market. Rather, it chose to reallocate the credit in the market by providing loans to institutions that participated in its new lending programs, while offsetting the effect of this lending on total credit through open market operations.

Why the Fed chose this unconventional approach is also unclear. Bernanke (2008) seems to suggest that the desire was not to increase the total liquidity in the economy but to provide liquidity to banks and other institutions that had illiquid, but sound, assets so that these institutions would continue to lend for productive purposes and avoid the inefficient liquidation of assets that were temporarily illiquid. It is also likely that the Fed was concerned that a significant increase in total liquidity might impair its ability to keep the federal funds rate close to the FOMC's target.[13]

Whatever the reason, it now appears that the Fed has abandoned the strategy of offsetting completely the effects of it new lending programs. Indeed, the Fed has injected historically large amounts of credit into the market. Such massive injections of base money have raised concerns about accelerating inflation. However, provided the increase is temporary and is removed once the need for additional liquidity is gone, as the Fed did in Y2K and 9/11, there is no reason that a temporary increase in base money should cause the long-term inflation rate to increase.

REFERENCES

Allen, Franklin and Gale, Douglas. *Understanding Financial Crises* (Clarendon Lectures in Finance). Oxford: Oxford University Press, 2007.

Bernanke, Ben S. "Liquidity Provision by the Federal Reserve." Presented at the Federal Reserve Bank of Atlanta Financial Markets Conference, Sea Island, Georgia, May 13, 2008; www.federalreserve.gov/newsevents/speech/bernanke20080513.htm.

Board of Governors of the Federal Reserve System. "Monetary Policy Report to the Congress."

[13] Keister, Martin, and McAndrews (2008) suggest that this is why the Fed chose not to increase the supply of total credit.

February 27, 2008; www.federalreserve.gov/boarddocs/hh/2008/february/fullreport.pdf.

Cecchetti, Stephen G. "Crisis and Responses: The Federal Reserve and the Financial Crisis of 2007-2008." NBER Working Paper No. 14134, National Bureau of Economic Research, June 2008; www.nber.org/ papers/w14134.pdf.

Friedman, Milton and Schwartz, Anna J. *A Monetary History of the United States, 1867-1960*. Princeton, NJ: Princeton University Press, 1963.

Keister, Todd; Martin, Antoine and McAndrews, James. "Divorcing Money from Monetary Policy." Federal Reserve Bank of New York *Economic Policy Review*, September 2008, *14*(2), pp. 41-56; http://www.ny.frb.org/research/EPR/08v14n2/0809keis.pdf.

Meltzer, Allan. *A History of the Federal Reserve, 1951-1986*. Volume 2A. Chicago: University of Chicago Press, 2009 (forthcoming).

Modigliani, Franco and Sutch, Richard C. "Innovations in Interest Rate Policy." *American Economic Review*, May 1966, *56*, pp. 178-97.

Neely, Christopher J. "The Federal Reserve Responds to Crises: September 11th Was Not the First." Federal Reserve Bank of St. Louis *Review*, March/April 2004, *86*(2), pp. 27-42; http://research.stlouisfed.org/publications/review/04/03/Neely.pdf.

Small, David H. and Clouse, James A. "The Scope of Monetary Policy Actions Authorized Under the Federal Reserve Act." *Topics in Macroeconomics*, 2005, *5*(1), Article 6.

Taylor, John B. and Williams, John C. "A Black Swan in the Money Market." Forthcoming in *American Economic Journal: Macroeconomics*, 2008a; preview at www.aeaweb.org/aej/mac/accepted/2008-0048.pdf.

Taylor, John B. and Williams, John C. "Further Results on a Black Swan in the Money Market." Stanford University, unpublished manuscript, 2008b; www.stanford.edu/~johntayl/Taylor-Williams-Further%20Results%20on%20Black%20Swan.pdf.

Thornton, Daniel L. "The Federal Reserve's Operating Procedures, Nonborrowed Reserves, Borrowed Reserves and the Liquidity Effect." *Journal of Banking and Finance*, 2001, *25*(9), pp. 1717-39.

Thornton, Daniel L. "Walter Bagehot, the Discount Window, and TAF." Federal Reserve Bank of St. Louis *Economic Synopses*, 2008, no. 27; http://research.stlouisfed.org/publications/es/08/ES0827.pdf.

Disallowances and Overcapitalization in the U.S. Electric Utility Industry

Stratford Douglas, Thomas A. Garrett, and Russell M. Rhine

Regulation of an industry often produces unintended consequences. Averch and Johnson (1962) argue that certain regulation of electric utilities provides utilities the incentive to purchase an inefficiently large amount of capital. Another possible and related unintended consequence of electric utility regulation is that regulatory cost disallowances on capital may also increase utilities' incentives to overcapitalize. The authors provide theoretical evidence that capital expenditure disallowances will increase the Averch and Johnson effect in some instances and thus may have contributed to the overcapitalization problem that regulation was designed to discourage. Our model shows that disallowances can reduce the rate of return on investment and thereby increase the Averch and Johnson distortion. (JEL D42, L43, L51)

Federal Reserve Bank of St. Louis *Review*, January/February 2009, *91*(1), pp. 23-31.

All households, firms, and government entities depend on one or more of the 3,170 electric utilities in the United States to provide a reliable source of energy. These electric companies consist of investor-owned, publicly owned, cooperative, and federal utilities.[1] Only 8 percent of these utilities are investor owned, but they produce approximately 75 percent of the total generating capability. Publicly owned and federal utilities each generate about 10 percent of the country's electricity, and cooperatives generate the remaining 4 percent. The revenue from retail sales (to ultimate consumers) for all electric utilities amounted to $326 billion in 2006 and represents about 2.5 percent of gross domestic product.

The industry increased output by 20 percent from 1995 to 2006, and generation capacity is expected to grow by another 8 percent over the next five years. Currently, nearly 98 percent of the existing capacity consists of fossil fuel power plants, nuclear reactors, hydroelectric power plants, and other renewable energy sources.[2] Although these sources all contribute to the total generation, fossil fuels generate the majority of electricity. Natural gas, coal, and petroleum supply 41 percent, 31 percent, and 6 percent of generation capabilities, respectively. Nuclear, hydroelectric, and other renewable sources comprise approximately 19 percent.[3] To date, the academic literature has devoted much attention to the U.S. electric

[1] Investor-owned utilities are private corporations that operate to produce a rate of return for their investors. Publicly owned utilities are nonprofit agencies owned by local governments. Cooperatives are owned by members of a community and typically operate in rural areas where investor-owned utilities are not economically feasible. Federal electric utilities are owned and operated by the federal government.

[2] Other renewable sources, as defined by the Energy Information Administration (2007), include wood, black liquor, other wood waste, municipal solid waste, landfill gas, sludge waste, tires, agriculture by-products, other biomass, geothermal, solar thermal, photovoltaic energy, and wind.

[3] For details, see Energy Information Administration (2007).

Stratford Douglas is an associate professor of economics at West Virginia University. Thomas A. Garrett is an assistant vice president and economist at the Federal Reserve Bank of St. Louis. Russell M. Rhine is an associate professor of economics at Saint Mary's College of Maryland. Lesli Ott provided research assistance.

utility industry. The primary reason for such interest is that electricity is used by all Americans, and firms in the industry enjoy a monopolistic market structure, at least at the distribution level. Although the academic literature is broad in scope, most articles fall into two categories. The first category is cost analysis—primarily the measuring of scale economies. That is, researchers attempt to determine where firms are operating on their long-run average cost curves and subsequently determine whether production costs can be lowered by having firms increase or decrease their scale of production. The second category, much larger than the first, is analysis of the regulatory aspect of the industry and the unanticipated consequences of those regulations. Relevant regulations involve not only those related to the environmental impact of electricity generation but also those regulating profits by setting the price that firms are allowed to charge for their electricity.

One specific issue that has sparked much attention is the overcapitalization of the electric utility industry—that is, electric utilities hold a quantity of capital that is greater than the cost-minimizing quantity. Averch and Johnson (1962; hereafter A-J) argued that privately owned utilities invest in capital beyond the cost-minimizing level in response to the incentives offered by regulation. The authors showed how a regulator, by tying a firm's allowed profit to its capital stock and offering a rate of return on capital that exceeds the marginal cost of capital, provides the firm the incentive to purchase an inefficiently large amount of capital. The A-J model has been thoroughly analyzed, discussed, and tested in the academic literature.[4]

[4] Nelson (1985) attributes the overcapitalization to utility overestimates of future demand growth in the late 1970s and early 1980s. Nemoto, Nakanishi, and Mandono (1993) find evidence of overcapitalization of electric utilities in Japan and attribute it to the A-J effect. Thompson, Islam, and Rose (1996) also find evidence of overcapitalization in the U.S. electric utility industry but they make no conclusions as to its source. Rungsuriyawiboon and Stefanou (2007) find evidence of the A-J effect using a dynamic duality model of intertemporal decisionmaking. Tests of the A-J effect in the electric utility industry using a production function rather than a cost function empirical specification have yielded conflicting results. Spann (1974) and Courville (1974) find evidence for the A-J effect using a translog and a Cobb-Douglas production function, respectively. Boyes (1976) uses a system of input demand functions derived from a production function and finds no evidence supporting the A-J theory.

The key aspect of the A-J model is that regulation of electric utilities results in the unintended consequence of overcapitalization. We argue in this paper that the A-J model is not complete and can thus be expanded to account for another aspect of electric utility regulation: A regulator's denial of cost recovery for some portion of the utility's capital expenditures provides an additional incentive for firms to overcapitalize. The basis for this incentive is that a portion of the firm's capital is excluded from the calculation of profit by the regulator. Firms will thus invest more in capital to maintain or increase profits.

One purpose of the regulatory disallowances was to make utility management accountable for cost overruns and thereby reduce their incentive to overcapitalize. Lyon and Mayo (2005) state that the disallowances were punitive and directed toward poorly managed firms. The disallowances succeeded in the sense that they apparently reduced utilities' appetite for constructing large new power plants; few large power plants and no new nuclear plants have been initiated in the past 20 years. But a subtler question remains as to whether the disallowances increased the efficiency of utilities' capital purchase decisions given the higher cost of capital.

We preface our theoretical framework by first providing an overview of the economic effects of regulation, including unintended consequences. We then provide a conceptual framework for overcapitalization that serves as a basis for our theoretical model, which expands the A-J model. We find theoretical support for the proposition that regulatory cost disallowances increase utilities' incentives to overcapitalize. Thus, the overcapitalization in the electric utility industry is a result not only of the A-J effect, but also of regulatory denial of cost recovery for a portion of a utility's capital expenditures.

THE ELECTRIC UTILITY INDUSTRY, MONOPOLY, AND REGULATION

The electric utility industry, like most public utilities, is considered a natural monopoly and has faced state and local regulations since the late

1800s.[5] As demonstrated in Figure 1, natural monopolies, by definition, exhibit decreasing average and marginal costs over a wide range of output because of high fixed costs (plants, equipment) and low variable and marginal costs.[6] Thus, one firm can produce most or all of the electricity demanded by consumers more cheaply than could multiple firms. Monopoly pricing involves charging a price (P_M) greater than the marginal cost (MC) of production and producing an output level (Q_M) less than that under perfect competition (Q_C), therefore resulting in a loss in economic efficiency.

The basic model of monopoly regulation posits that regulators aim to reduce the price charged by the monopolist and expand the monopolist's output. One common approach is to set the price equal to the marginal cost of production (point E). This mimics pricing under perfect competition. However, given the cost structure of monopoly, marginal costs are below average costs (AC), so marginal cost pricing often results in a financial loss for the monopolist. Average cost pricing deviates from the competitive price and output level (because average costs > marginal costs), but still results in a price and output level (point F) that approximates the competitive solution.

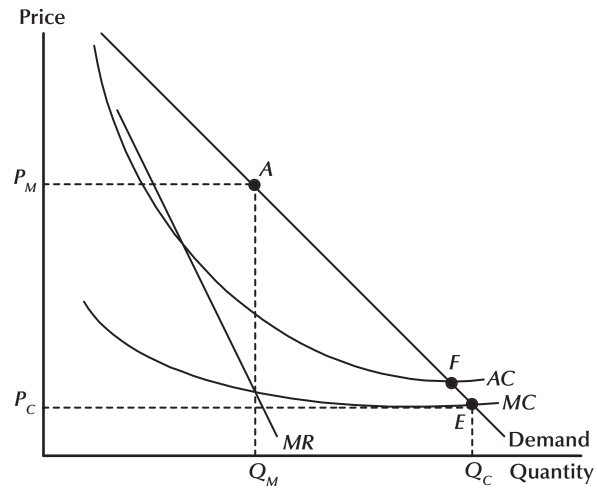
Although in theory the regulation of monopoly pricing is fairly straightforward, in reality it may be difficult to achieve the price and output levels that would exist under perfect competition, given that regulation occurs in political markets. As first discussed by Stigler (1971), consumers and producers have different objectives with regard to monopoly prices—consumers prefer lower prices and greater output, whereas the monopolist prefers higher prices and lower output. According to Stigler, it is reasonable to assume that both groups exert political pressure to set regulatory outcomes in their favor.

Stigler's model shows that a vote-maximizing regulator will set a utility's sale price of electricity such that the marginal gain in support from producers is just offset by the loss in consumer votes. Thus, in Figure 1, political competition between consumers and producers will ensure that the regulated price will lie somewhere between P_M and P_C and the regulated output will lie somewhere between Q_M and Q_C. The exact location will depend on the relative strengths of consumer and producer groups in exerting political pressure. Given that the cost of organizing producer groups is much less than the cost of organizing thousands or millions of consumers (Olson, 1965; Peltzman, 1976), producers are likely to exert more political pressure than are consumers; as a result, regulation will likely favor producers.

Another aspect of regulation that has garnered attention in the literature—and is most relevant for the purpose of our paper—is the potential for unintended consequences as a result of regulation. Unintended consequences are, as the term suggests, unanticipated effects from policy actions. Examples from other industries include those described by Hall, Propper, and Reenan (2008), who find that regulated pay for medical staff across geographically heterogeneous labor markets results in problems with recruiting, retaining, and motivating high-quality workers, which ultimately

Figure 1

Structure of Natural Monopoly

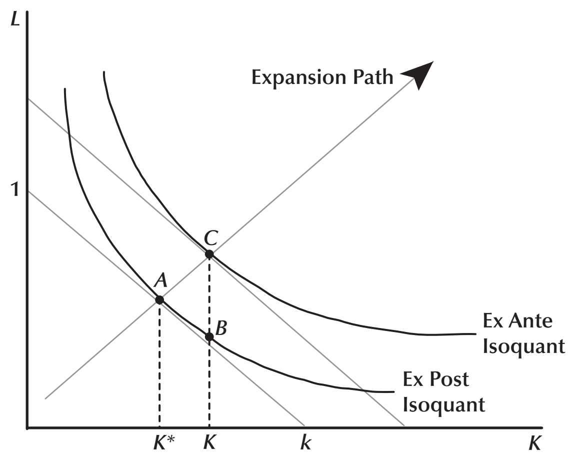


[5] See Warkentin-Glenn (2006) for a history of the electric utility industry.

[6] The cost curves shown in Figure 1 are for a representative natural monopolist and may not represent the identical cost and pricing structure of all monopolists.

Figure 2

Overcapitalization



affects hospital performance; and Nelson (2003), who finds that the regulation of alcohol advertising initiated to restrict demand for one type of product creates increased demand for other alcoholic products.

As mentioned earlier, overcapitalization in the electric utility industry is one unintended consequence of regulating the industry. Specifically, a firm is motivated to purchase an inefficiently large amount of capital because a regulator ties the firm's allowed profit to its capital stock. We argue that overcapitalization will be exacerbated when a portion of a firm's capital is excluded from the calculation of profit by the regulator, and firms will thus invest more in the type of capital not excluded from the profit calculation to maintain or increase profits.

Overcapitalization Explained

Many industry observers, both inside and outside academia, believe that the electric utility industry in the United States is overcapitalized. Our theoretical model presented in the next section focuses on overcapitalization, and so we discuss overcapitalization in general to better lay the foundation for our theoretical model.

In Figure 2, K^* is the cost-minimizing capital stock for production along the ex post isoquant. Suppose we observe a firm producing inefficiently at point *B* using capital stock, *K*. The A-J model explains a firm's decision to produce at point *B* as a rational response to regulatory incentives. Once it arrives at *B*, it will tend to stay there. More intuitively, the price that an electric utility is allowed to charge for its output is set by industry regulators. The regulators choose a price so that the firm will generate enough revenue to cover the operating costs, which include labor expenses, and to generate enough profit to pay the investors a fair rate of return on capital. Thus, with production inputs of capital and labor, the firm knows that an increase in labor expenses will be offset by an increase in revenues by the same amount. However, an increase in capital investment will be offset by an increase in revenue that is greater than the increased costs to ensure that investors receive a fair rate of return. Thus, the firm has an incentive to increase its investment in capital. Our theoretical exercise in the next section of the paper determines whether a firm will choose point *B* instead of *A* because of regulatory incentives or for some other reason.

Overcapitalization can appear in many different guises. Utilities that need new capacity might avoid leasing available power plants, preferring to build their own. They might resist selling power plants that they do not need. They might resist joining efficient power pools. They might inefficiently choose capital-intensive expenditures for pollution abatement (such as scrubbers) over less capital-intensive alternatives (such as co-firing or "green power" purchases). They also might appear overeager to move to underground distribution systems. They might impose excessive safety and reliability standards on themselves, resulting in a "reserve margin" in the form of an idle generation plant that exceeds any rational requirements. As Kahn (1988) explains, utilities may have "excessively high (because excessively costly) standards of reliability and uninterruptibility of service, with correspondingly high and costly specification for the equipment they employ." Finally, utilities may allow themselves to be overcharged for capital equipment, as occurred in the electrical equipment manufacturers' price conspiracy of the 1950s (Emery, 1973).

Overcapitalization also can occur because of forecast error when the firm overestimates future demand. Faulty forecasting is a reasonable suspect as a cause of utility overcapitalization. The late 1970s and early 1980s saw both a decrease in the U.S. rate of economic growth and, because of the increase in relative energy prices, the first significant divergence between rates of growth in the economy and in the demand for energy. Traditional methods of demand forecasting failed to account for these trends, and utilities continued to build large power plants to serve demand that, in many cases, never materialized. Exacerbating this trend was the greatly increased cost of nuclear power as a result of the regulatory response to the Three Mile Island incident in 1979.

Large generation plants must be built well before anticipated demand is realized. If the ex ante isoquant in Figure 2 were the anticipated level of demand, then *K* would be the ex ante cost-minimizing level of capital. However, when actual demand is realized ex post, the firm finds itself operating off its least-cost expansion path at point *B*. If presented with efficient incentives, the firm would move to point *A* as quickly as possible. Thus, if overcapitalization arose solely in response to overestimation of demand, then overcapitalization should decline over time as the firm adapts to the unexpectedly low-demand environment.

Past studies on adapting to a world in which regulators occasionally deny recovery or return on part or all of a utility's capital investment have shown varying results. Lyon (1991), Gal-Or and Spiro (1992), and Gilbert and Newbery (1994) argue that capital disallowances may decrease investment, whereas Tiesberg (1993) finds that capital disallowances may increase investment. The disallowance might be expected to increase the riskiness of the firm's investments, and this increased riskiness of capital could cause utilities to reduce their use of capital. If so, the model presented in this paper identifies a countervailing incentive—a tendency to overcapitalize even further because of the disallowance effect on a utility's de facto allowed rate of return.

The A-J Model

The A-J model depicts a profit-maximizing firm that is subject to a regulated rate of return. The production function specification may be inappropriate for the traditional electric utility, given the generally accepted stylized facts of the industry. A production function assumes that a firm maximizes output for a given level of inputs; however, a traditional electric utility is more likely to find its output level fixed because its price is regulated within its geographically defined service area. It will choose its input quantities endogenously in response to largely exogenous factor market prices. Therefore, the dual cost function, in which input quantities are chosen in response to fixed factor prices and output levels, is a better model for the behavior of the electric utility.

A-J demonstrate that if the allowed rate of return exceeds the cost of capital, then a firm's capital stock will increase to the point where the cost of capital exceeds its marginal product. Let $R(Y)$, P, Y, P_i, K, L, and s_K be the revenue function, output price, the quantity of output, the price of the inputs, the quantity of capital, the quantity of labor, and the allowed rate of return, respectively, where $i = K, L$. The price of capital, P_K, is the interest cost in holding plant and equipment, which differs from the acquisition cost of capital, c. We assume that $c = 1$; thus, the regulator constrains the firm to an allowed rate of return, s_K, that will cover the physical quantity of capital, $c \cdot K$, after operating expenses are deducted from revenue. That is,

$$\frac{PY - P_L L}{K} \leq s_{K^*}.$$

Equation (1) below is the Lagrangian for a profit-maximizing firm. The firm maximizes profits subject to two constraints. The first constraint is the regulatory-constrained allowed rate of return, and the second constraint is output that is subject to the production function, $Y \leq F(K, L)$. Equations (2) through (6) are the first-order conditions that result from maximizing equation (1):

$$(1) \quad \begin{aligned} \mathcal{L}(K, L, \lambda, \mu) &= R(Y) - P_K K - P_L L \\ &+ \lambda\left(s_K K - R(Y) + P_L L\right) + \mu\left(F(L, K) - Y\right) \end{aligned}$$

$$(2)\qquad \frac{\delta \mathcal{L}}{\delta L} = P_L(\lambda - 1) + \mu F_L = 0$$

$$(3)\qquad \frac{\delta \mathcal{L}}{\delta \mu} = F(K,\, L) - Y_L = 0$$

$$(4)\qquad \frac{\delta \mathcal{L}}{\delta K} = -P_K + \lambda s_K + \mu F_K = 0$$

$$(5)\qquad \frac{\delta \mathcal{L}}{\delta Y} = (1-\lambda)R' - \mu = 0$$

$$(6)\qquad \frac{\delta \mathcal{L}}{\delta \lambda} = s_K K - R + P_L L = 0.$$

We assume that $R' > 0$, $F_L > 0$, $F_K > 0$ and that the firm earns a profit (i.e., $s_K > P_K$). We also set $G = R(\mathrm{F}(K,L))$ and assume that it is concave so that the second-order condition for maximization is satisfied. Therefore, the marginal revenue products of labor and capital are $G_L = R'F_L$ and $G_K = R'F_K$, respectively. Finally, we assert that $1 > \lambda > 0$ and $\mu > 0$ (Takayama, 1993). By combining equations (2) and (5), it follows that the firm uses labor efficiently: G_L equals the wage; but by combining equations (4) and (5), it follows that firms do not use capital efficiently: G_K is less than P_K, as shown in equation (7):

$$(7)\qquad G_L = P_L, \quad G_K < P_K.$$

By totally differentiating equation (6) with respect to s_K and applying the efficient use of labor from equation (7), the A-J effect can be produced and is shown in equation (8).[7] Decreasing the allowed rate of return results in the firm increasing its capital stock:

$$(8)\qquad \begin{aligned} & K + s_K \frac{dK}{ds_K} - G_K \frac{dK}{ds_K} + P_L \frac{dL}{ds_K} - G_L \frac{dL}{ds_K} = 0 \\ & \Rightarrow K = (G_K - s_K)\frac{dK}{ds_K} + (G_L - P_L)\frac{dL}{ds_K} \\ & \Rightarrow \frac{dK}{ds_K} = \frac{K}{G_K - s_K} < 0. \end{aligned}$$

[7] See Takayama (1993, pp. 215-16) for a more detailed discussion and rigorous proof.

The capital stock, K, plays a twofold role in a regulated firm, as both a productive input and the rate base for computing allowed profit. The rate base is defined as the amount of capital expenditures that the regulator uses to determine the costs that should be passed on to consumers. The regulator determines whether expenditures on capital may be included in the firm's rate base and therefore paid for by utility customers. If a portion of a firm's capital expenditure was not prudently incurred, then the regulator may disallow that portion (i.e., exclude it from the rate base). Many utilities incurred massive disallowances in the 1980s, primarily because of cost overruns and tighter safety standards for nuclear power plants.

We suggest that the disallowances in fact exacerbated the A-J effect in some instances, thereby reducing the efficiency of capital use. The exact effect depends on whether the disallowed power plant is capable of producing electricity for sale. Suppose that the regulator disallows a capital expenditure on a power plant that represents some proportion, δ, of the utility's total rate base, where $0 < \delta < 1$. If the power plant is never completed or is not allowed to operate, then the full capital expenditure appears as a cost, but only the allowed portion appears in the rate base and the production function.[8] Thus, the disallowed capital expenditure is useless—it is not generating electricity or revenues for the firm. The Lagrangian for disallowed useless capital is shown by equation (9):

$$(9)\qquad \begin{aligned} & \mathcal{L}_{(useless)}(K,\, L,\, \lambda,\, \mu) = R(Y) - P_K K - P_L L \\ & +\lambda\big(s_K(1-\delta)K - R(Y) + P_L L\big) \\ & +\mu\big(F\big((1-\delta)K,\, L\big) - Y\big), \end{aligned}$$

where $R(Y) \equiv G\big((1-\delta)K,\, L\big).$

The firm incurs the full cost of purchasing K, but the allowed rate of return, s_K, applies only to the

[8] An example of a completed power plant that was unable to be used for production is the Shoreham nuclear power plant in New York; on completion, it was never allowed to operate because of safety concerns.

rate base $(1-\delta)K$ and only $(1-\delta)K$ is productive. Essentially, the disallowance reduces the firm's capital stock, both as productive input and as a portion of the rate base, while retaining a sunk cost of $P_K\,\delta K$ on the firm. Equations (10) through (14) are the first-order conditions that result from the maximization of equation (9):

$$(10)\quad \frac{\delta \mathcal{L}}{\delta L} = -P_L(\lambda-1)+\mu F_L = 0$$

$$(11)\quad \frac{\delta \mathcal{L}}{\delta \mu} = F\big((1-\delta)K,\ L\big)-Y = 0$$

$$(12)\quad \frac{\delta \mathcal{L}}{\delta K} = -P_K+\lambda(1-\delta)s_K+(1-\delta)\mu F_K = 0$$

$$(13)\quad \frac{\delta \mathcal{L}}{\delta Y} = (1-\lambda)R'-\mu = 0$$

$$(14)\quad \frac{\delta \mathcal{L}}{\delta \lambda} = s_K(1-\delta)K-R+P_L L = 0.$$

The same first- and second-order conditions as previously stated hold, and the firm earns a profit, $(1-\delta)s_K > P_K$. Combining equations (10) and (13) indicates efficient use of labor, as shown in equation (15):

$$(15)\quad G_L = P_L.$$

The incremental input distortion caused by a change in the allowed rate of return, s_K, is the same in the model with a useless capital disallowance as it is in the classic A-J model. This can be seen by totally differentiating equation (14) with respect to s_K and applying equation (15). Equation (16) shows the effect of a change in s_K on the capital stock, K:

(16)

$$(1-\delta)K+(1-\delta)s_K\frac{dK}{ds_K}-(1-\delta)G_K\frac{dK}{ds_K}$$
$$+P_L\frac{dL}{ds_K}-G_L\frac{dL}{ds_K}=0$$
$$\Rightarrow (1-\delta)K=(G_K-s_K)(1-\delta)\frac{dK}{ds_K}+(G_L-P_L)\frac{dL}{ds_K}$$
$$\Rightarrow \frac{dK}{ds_K}=\frac{K}{G_K-s_K}<0.$$

Although no change occurs in the marginal effect of an increase in the allowed rate of return, an increase in the disallowance of useless capital does, in fact, increase the amount of capital purchased. Equation (17) is the result of totally differentiating equation (14) with respect to δ, and it indicates that an increase in the scale of the disallowance causes an increase in K:

$$(17)\quad \frac{dK}{d\delta}=\frac{s_K K}{(1-\delta)(s_K-G_K)}>0.$$

More commonly, the disallowed capital is capable of producing electricity (i.e., it is "useful"), but it is either not needed to serve the utility's captive customer base or its construction costs are judged excessive. In the latter case, the utility will take the disallowed costs out of its rate base, but it may continue to operate and sell the power either to its customers or off-system on the wholesale market. Thus, as shown in equation (18), the disallowed unit remains in the production function

$$(18)\quad \begin{aligned}&\mathcal{L}_{useful}(K,\ L,\ \lambda,\ \mu)=R(Y)-P_K K-P_L L\\&+\lambda\big(s_K(1-\delta)K-R(Y)+P_L L\big)\\&+\mu\big(F(K,\ L)-Y\big),\end{aligned}$$

where $R(Y)\equiv G(K,\ L)$.

Because the first-order conditions with respect to Y and L are identical to that of $\mathcal{L}_{useless}$, $G_L = P_L$ still holds. However, the remaining first-order conditions do differ and are shown as equations (11′), (12′), and (14′):

$$(11')\quad \frac{\delta \mathcal{L}}{\delta \mu}=F(K,\ L)-Y=0$$

$$(12')\quad \frac{\delta \mathcal{L}}{\delta K}=-P_K+\lambda(1-\delta)s_K+\mu F_K=0$$

$$(14')\quad \frac{\delta \mathcal{L}}{\delta \lambda}=s_K(1-\delta)K-R+P_L L=0.$$

The effect of changing the allowed rate of return, s_K, is somewhat different in the presence of useful but disallowed capital,

$$(19)\qquad \frac{dK}{ds_K}=\frac{(1-\delta)K}{G_K-(1-\delta)s_K}<0.$$

The inequality in equation (19) holds if the post-disallowance rate of return exceeds the marginal revenue product of capital: that is, $(1-\delta)s_K > G_K$. Note that if there were no disallowed capital in equation 19, the result would be identical to the A-J model shown by equation (8). However, equation (20) indicates that the A-J distortion effect of regulatory tightening for disallowed *useful* capital exceeds the distortion for a firm with disallowed *useless* capital:

$$(20)\qquad \frac{dK}{ds_{K(useful)}}=\left[\frac{(1-\delta)(s_K-G_K)}{(1-\delta)s_K-G_K}\right]\frac{dK}{ds_{K(useless)}}$$
$$\text{and}\quad \frac{(1-\delta)(s_K-G_K)}{(1-\delta)s_K-G_K}>1.$$

In addition, as equation (21) shows, an increase in the disallowance of useful capital has a positive marginal distortive effect on the capital stock:

$$(21)\qquad \frac{dK}{d\delta}=\frac{s_K K}{(1-\delta)s_K-G_K}>0.$$

The marginal distortive effect of increasing δ for useful capital (equation (21)), exceeds that of useless capital (equation (17)), by the same factor as in equation (20): that is, $[(1-\delta)(s_K-G_K)]/[(1-\delta)s_K-G_K]$.

Thus, the modified A-J model predicts that a firm subject to a capital disallowance will overcapitalize to a greater degree than a firm without a disallowance. More intuitively, our results indicate that because a utility's profit is equal to $[(1-\delta)s_K-P_K]K$, the existence of disallowed capital, δ, will cause the utility's profit to decrease, all other things equal. For the firm to regain those lost profits, either its capital stock, K, or the allowed rate of return, s_K, must increase. Consequently, the only way to regain the profits is by investing in capital because its allowed rate of return cannot be changed by the firm, as it is set by the regulators.

In addition, a firm with disallowed capital that is operable will experience a greater A-J distortion than a firm with disallowed capital that is not operable or a firm with no disallowed capital. This difference is due to the fact that the operable disallowed capital continues to generate revenue for the firm; subsequently, the marginal revenue product of capital is not reduced by the existence of capital that generates no revenue.

CONCLUSION

We provide theoretical evidence that capital expenditure disallowances reinforce the Averch-Johnson effect and thereby may have contributed to the overcapitalization problem they were designed to decrease. The theoretical model shows that disallowances, especially of useful plants, reduce the de facto allowed rate of return on capital and thereby increase the Averch-Johnson input distortion. All these results support the idea that overcapitalization is a profit-maximizing response to rate-of-return regulatory incentives and not a mistake caused by incorrect demand forecasting. These results have some relevance to current policy. Regulators and regulated firms in several industries, including natural gas, electricity, and telecommunications, are taking capital assets out of their rate bases. Currently, the reason for such action is not imprudence but deregulation. Our results suggest that the way in which this process occurs matters. In particular, taking the assets out of the rate base, but leaving them in the hands of a regulated firm that is subject to an overall rate-of-return constraint, could result in inefficient overcapitalization. In general, regulators need to be aware that their policies could have unanticipated consequences.

REFERENCES

Averch, Harvey and Johnson, Leland J. "Behavior of the Firm under Regulatory Constraint." *American Economic Review*, December 1962, *52*(5), pp. 1052-69.

Boyes, William J. "An Empirical Estimation of the Averch-Johnson Effect." *Economic Inquiry*, March 1976, *14*(1), pp. 25-35.

Courville, Leon. "Regulation and Efficiency in the Electric Utility Industry." *Bell Journal of Economic and Management Science*, Spring 1974, *5*(1), pp. 53-74.

Emery, David E. "Regulated Utilities and Equipment Manufacturers' Conspiracies in the Electrical Power Industry." *RAND Journal of Economics*, Spring 1973, *4*(1), pp. 322-37.

Energy Information Administration. "Electric Power Annual, 2006," released October 22, 2007; www.eia.doe.gov/cneaf/electricity/epa/epa_sum.html.

Gal-Or, Esther and Spiro, Michael H. "Regulatory Regimes in the Electric Power Industry: Implications for Capacity." *Journal of Regulatory Economics*, September 1992, *4*(3), pp. 263-78.

Gilbert, Richard J. and Newbery, David M. "The Dynamic Effect of Regulatory Constitutions." *RAND Journal of Economics*, Winter 1994, *25*(4), pp. 538-54.

Hall, Emma P.; Propper, Carol and van Reenen, John. "Can Pay Regulation Kill? Panel Data Evidence on the Effect of Labor Markets on Hospital Performance." NBER Working Paper No. 13776, National Bureau of Economic Research, February 2008; www.nber.org/papers/w13776.

Kahn, Alfred E. *The Economics of Regulation: Principles and Institutions*. Volume 2. Cambridge, MA: MIT Press, 1988.

Lyon, Thomas P. "Regulation with 20-20 Hindsight: 'Heads I Win, Tails You Lose?'" *RAND Journal of Economics*, Winter 1991, *22*(4), pp. 581-95.

Lyon, Thomas P. and Mayo, John W. "Regulatory Opportunism and Investment Behavior: Evidence from the U.S. Electric Utility Industry." *RAND Journal of Economics*, Autumn 2005, *36*(3), pp. 628-44.

Nelson, Jon P. "Advertising Bans, Monopoly, and Alcohol Demand: Testing for Substitution Effects Using State Panel Data." *Review of Industrial Organization*, February 2003, *22*(1), pp. 1-25.

Nelson, Randy A. "Returns to Scale from Variable and Total Cost Functions: Evidence from the Electric Power Industry." *Economic Letters*, 1985, *18*(2-3), pp. 271-76.

Nemoto, Jiro; Nakanishi, Yasuo and Mandono, Seishi. "Scale Economies and Over-Capitalization in Japanese Electric Utilities." *International Economic Review*, May 1993, *34*(2), pp. 431-40.

Olson, Mancur. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press, 1965.

Peltzman, Sam. "Toward a More General Theory of Regulation." *Journal of Law and Economics*, August 1976, *19*(2), pp. 211-40.

Rungsuriyawiboon, Supawat and Stefanou, Spiro E. "Dynamic Efficiency Estimation: An Application to U.S. Electric Utilities." *Journal of Business and Economic Statistics*, April 2007, *25*(2), pp. 226-38.

Spann, Robert M. "Rate of Return Regulation and Efficiency in Production: An Empirical Test of the Averch-Johnson Thesis." *Bell Journal of Economic and Management Science*, Spring 1974, *5*(1), pp. 38-52.

Stigler, J. George. "The Theory of Regulation." *Bell Journal of Economics and Management Science*, Spring 1971, *2*(1), pp. 3-21.

Takayama, Akira. *Analytical Methods in Economics*. Ann Arbor: University of Michigan Press, 1993.

Teisberg, Elizabeth Olmstead. "Capital Investment Strategies under Uncertain Regulations." *RAND Journal of Economics*, Winter 1993, *24*(4), pp. 591-604.

Thompson, Herbert G.; Hovde, David A.; Irwin, Luis; Islam, Mufakharul and Rose, Kenneth. "Economies of Scale and Vertical Integration in the Investor-Owned Electric Utility Industry." NRRI Report No. 96-05, National Regulatory Research Institute, January 1996; http://nrri.org/pubs/electricity/96-05.pdf.

Warkentin-Glenn, Denise. *Electric Power Industry: In Nontechnical Language*. Second Edition. Tulsa, OK: PennWell Corporation, 2006.

Optimal Response to a Transitory Demographic Shock in Social Security Financing

Juan C. Conesa and Carlos Garriga

The authors consider a transitory demographic shock that affects negatively the financing of retirement pensions—that is, workers either would have to pay more or retirees would receive less. In contrast to the existing literature, the authors endogenously determine optimal policies rather than explore the implications of exogenous parametric responses. Their approach identifies optimal strategies of the Social Security Administration to guarantee the financial sustainability of existing retirement pensions in a Pareto-improving way. Hence, no cohort will pay the cost of the demographic shock. The authors find that the optimal strategy is based on the following ingredients: elimination of compulsory retirement, a change in the structure of labor income taxation, and a temporary increase in the level of government debt. (JEL D58, D91, H55)

Federal Reserve Bank of St. Louis *Review*, January/February 2009, *91*(1), pp. 33-48.

The financial sustainability of the Social Security system is an important policy concern due to the aging of the U.S. population and in particular of the baby-boom generation. According to estimates of the Social Security Administration, the dependency ratio (measured as population age 65 or older over population between ages 20 and 64) will increase from its present 21 percent to 27 percent in the year 2020, 37 percent in 2050, and 42 percent in 2080 under the scenario they call the medium population growth (Figure 1).

Under this demographic scenario, the Social Security system, which is a pay-as-you-go (PAYG) program, will face clear financial imbalances unless some reforms are introduced. In this paper, we explore the optimal response to a transitory demographic shock that affects negatively the financing of retirement pensions.[1] In contrast to existing literature, we follow an approach that is similar to that used in Conesa and Garriga (2008) and endogenously determine optimal policies rather than exploring implications of exogenous parametric policies. Our approach determines the optimal strategy of the Social Security Administration to guarantee the financial sustainability of current retirement pensions in the least distortionary way. Moreover, no cohort will have to pay the welfare cost of the demographic shock.

Notice that we are concerned only about efficiency considerations in the financing of retirement pensions rather than about the efficiency of their existence in the first place. Their existence

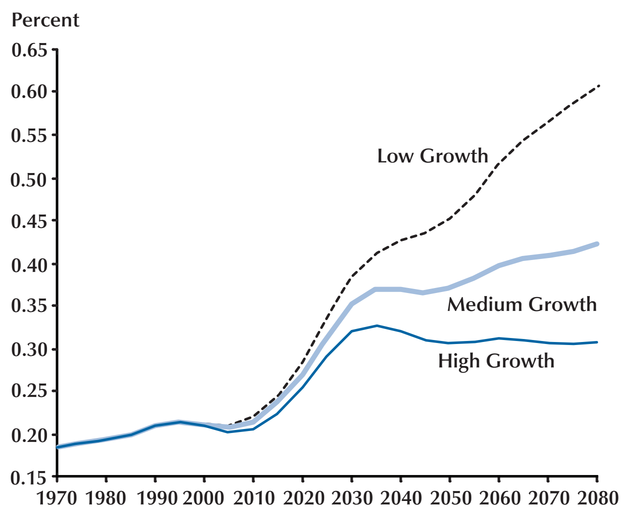
[1] In our artificial economy, we assume the transitory nature of the demographic shock for computational convenience, while Figure 1 clearly shows the permanent nature of the future demographic shock faced by the U.S. population structure.

Juan C. Conesa is an associate professor at the Universitat Autònoma de Barcelona. Carlos Garriga is an economist at the Federal Reserve Bank of St. Louis. The authors thank Mark Keightley, Georges de Ménil, Pierre Pestieau, and two anonymous referees for their very useful comments. Both authors acknowledge financial support from the Spanish Ministerio de Educación y Ciencia, SEJ2006-03879. Conesa also acknowledges support from the Barcelona Economics Program and Generalitat de Catalunya through 2005SGR00447.

This article was originally published in *Pension Strategies in Europe and the United States* (Robert Fenge, Georges de Ménil, and Pierre Pestieau, editors), Cambridge, MA: MIT Press, pp. 98-113. Reprinted here with permission.

Figure 1

Population Age 65 and Older Over Population Ages 20 to 64



SOURCE: Social Security Administration.

might be justified on different grounds.[2] We do not model why Social Security was implemented in the first place or why Social Security benefits are provided through a potentially inefficient tax system.

We consider for our experiments an unexpected transitory demographic shock, even though these shocks are certainly predictable by looking at Figure 1. If the demographic shock is predictable, the fiscal authority should have reacted to it in advance. However, we believe it is more interesting to focus on what should be done from now on rather than on what should have been done. In this sense, prediction of a demographic shock without action is equivalent to the shock being unexpected. However, the transitory nature of the shock considered is a limitation of the analysis driven by computational tractability.

The quantitative evaluation of Social Security reforms has been widely analyzed in the literature.[3] Demographic considerations play an important role in the Social Security debate, but there are few quantitative studies of policy responses to demographic shocks and none to our knowledge from an optimal fiscal policy perspective. In particular, De Nardi, Imrohoroglu, and Sargent (1999) consider the economic consequences of different alternative fiscal-adjustment packages to solve the future Social Security imbalances associated with the projected demographics in the United States. They find that all fiscal adjustments impose welfare losses on transitional generations. In particular, policies that partially reduce retirement benefits (by taxing benefits, postponing retirement, or taxing consumption), or that gradually phase benefits out without compensation yield welfare gains for future generations but make most of the current generations worse off. They conclude that a sustainable Social Security reform requires reducing distortions in labor supply or in consumption and saving choices and some transition policies to compensate current generations (issuing government debt). Our approach allows for the endogenous determination of such policies in a way that nobody faces welfare losses. In other words, everybody will be guaranteed the same level of welfare as in the benchmark economy without a demographic shock. However, for computational tractability we will substantially simplify the nature of the demographic shocks relative to De Nardi, Imrohoroglu, and Sargent (1999).

Jeske (2003) also analyzes payroll adjustments to demographic shocks in an economy similar to ours. He finds that in contrast with the benchmark economy not all cohorts are worse off due to the arrival of the baby boomers. The parents of the baby boomers gain about 0.5 percent of average lifetime consumption, the baby boomers lose 1 percent, the children of boomers gain 2 percent, and the grandchildren lose more than 2 percent. The intuition for this result comes from move-

[2] One basic reason could be dynamic inefficiencies (see Diamond, 1965, or Gale, 1973). Also, even in a dynamically efficient economy, Social Security might be sustained because of political economy considerations (see Grossman and Helpman, 1998; Cooley and Soares, 1999; or Boldrin and Rustichini, 2000). Also, Social Security might be part of some general social contract, as in Boldrin and Montes (2005).

[3] Feldstein and Liebman (2001) summarize the discussion on transition to investment-based systems, analyzing the welfare effects and the risks associated with such systems.

ments in factor prices implied by the demographic shock and the implied payroll taxes adjustment to balance the per-period government budget constraint.

In contrast to both of them, we do not analyze the different implications of exogenously specified strategies to guarantee sustainability but instead optimize over this policy response to demographic shocks following the Ramsey approach. The quantitative analysis of optimal fiscal policy in overlapping generations economies was pioneered by Escolano (1992) and has been recently considered by Erosa and Gervais (2002) and Garriga (1999). Conesa and Garriga (2008) used a similar framework to analyze the design of Social Security reforms, and therefore the focus was on efficiency considerations, abstracting from sustainability issues.

Our main conclusions indicate that the optimal strategy in absorbing a negative demographic shock consists of the following:

- Changing the age structure of labor-income taxation. In particular, labor-income taxes of the young should be substantially decreased.
- Eliminating compulsory retirement and allowing cohorts older than age 65 to supply labor in the market.
- Increasing the level of government debt during the duration of the demographic shock and then repaying it slowly.

We find that the welfare gains will be concentrated for generations born in the distant future after the demographic shock is over, while it does maintain the benchmark welfare level for existing cohorts and current newborns during the shock. Therefore, no generation is worse off along the fiscal-adjustment process implied by the demographic shock. This result contrasts with the findings of De Nardi, Imrohoroglu, and Sargent (1999), and Jeske (2003), where either current or future generations suffer important welfare losses. More important, we find that a sustainable Social Security reform does not necessarily require reducing distortions in consumption and saving choices. A reduction in labor supply distortions and the issuing of government debt are sufficient to compensate current generations.

In addition, we show that the welfare costs of distortionary taxation are quantitatively important right after the demographic shock but are relatively less important in the long run.

The distortionary impact of the financing of pensions in our artificial economy is assumed rather than endogenously determined. In our benchmark economy, pensions are financed through linear age-independent payroll taxes, and individuals do not establish a link between their individual contributions to the system and their future pensions. Hence, all the welfare gains obtained in our analysis are generated by the minimization of distortions and the redistribution of these additional resources. Indeed, it could not be otherwise since the possibility of Pareto improvements exists only because of the presence of distortions.

We also show that when the income from retirement pensions is not taxable, the government could use this fact to replicate lump-sum taxation and achieve first-best allocations. Yet since we want to focus on an environment where the government is restricted to distortionary taxation, we consider only an environment where the fiscal treatment of retirement pensions is constrained to be the same as that of regular labor income.

The rest of the paper describes the benchmark theoretical framework used, our method of parameterizing our benchmark economy, the optimal fiscal policy problem using the primal approach, the experiment we perform, the demographic shock, and our analysis of the optimal response.

THE THEORETICAL ENVIRONMENT IN THE BENCHMARK ECONOMY

Households

The economy is populated by a constant measure of households who live for I periods. These households are forced to retire in period i_r. We denote by $\mu_{i,t}$ the measure of households of age i in period t. Preferences of a household born in period t depend on the stream of consumption and leisure this household will enjoy. Thus, the utility function is given by

$$(1)\qquad U\left(c^t, l^t\right) = \sum_{i=1}^{I} \beta^{i-1} u\left(c_{i,t+i-1}, 1 - l_{i,t+i-1}\right).$$

Every period, each household owns one unit of time that they can allocate to work or leisure. One unit of time devoted to work by a household of age i translates into ε_i efficiency units of labor in the market, and these are constant over time.

Technology

The production possibility frontier is given by an aggregate production function $Y_t = F(K_t, L_t)$, where K_t denotes the capital stock at period t, and

$$L_t = \sum_{i=1}^{I} \mu_{i,t} \varepsilon_i l_{i,t}$$

is the aggregate labor endowment measured in efficiency units. We assume the function F displays constant returns to scale, is monotonically increasing, is strictly concave, and satisfies the Inada conditions. The capital stock depreciates at a constant rate δ.

Government

The government influences this economy through the Social Security and the general budget. For simplicity, we assume that initially (before the demographic shock) these two programs operate with different budgets. Pensions (p_t) are financed through a payroll tax (τ_t^p), and the Social Security budget is balanced. On the other hand, the government collects consumption taxes (τ_t^c), labor income taxes (τ_t^l), and capital-income taxes (τ_t^k) and issues public debt (b_t) to finance an exogenously given stream of government consumption (g_t).

Thus, the Social Security and government budget constraints are respectively given by

$$(2)\qquad \tau_t^p w_t \sum_{i=1}^{i_r-1} \mu_i \varepsilon_i l_{i,t} = p_t \sum_{i=i_r}^{I} \mu_i, \text{ and}$$

$$(3)\qquad \begin{aligned} &\tau_t^c \sum_{i=1}^{I} \mu_i c_{i,t} + \tau_t^l \left(1 - \tau_t^p\right) w_t \sum_{i=1}^{i_r-1} \mu_i \varepsilon_i l_{i,t} \\ &+ \tau_t^k r_t \sum_{i=1}^{I} \mu_i a_{i,t} + b_{t+1} = g_t + \left(1 + r_t\right) b_t. \end{aligned}$$

In response to the demographic shock, however, both budgets are integrated, and we allow the government to transfer resources across budgets to finance the retirement pensions.

Market Arrangements

We assume there is a single representative firm that operates the aggregate technology, taking factor prices as given. Households sell an endogenously chosen fraction of their time as labor ($l_{i,t}$) in exchange for a competitive wage of w_t per efficiency unit of labor. They rent their assets ($a_{i,t}$) to firms or the government in exchange for a competitive factor price (r_t) and decide how much to consume and save out of their disposable income. The sequential budget constraint for a working-age household is given by

$$(4)\qquad \begin{aligned} &\left(1 + \tau_t^c\right) c_{i,t} + a_{i+1,t+1} = \left(1 - \tau_t^l\right)\left(1 - \tau_t^p\right) w_t \varepsilon_i l_{i,t} \\ &+ \left(1 + \left(1 - \tau_t^k\right) r_t\right) a_{i,t}, \quad i = 1, \ldots, i_r - 1. \end{aligned}$$

On retirement, households do not work and receive a pension in a lump-sum fashion. Their budget constraint is

$$(5)\qquad \begin{aligned} &\left(1 + \tau_t^c\right) c_{i,t} + a_{i+1,t+1} = \left(1 - \tau_t^l\right) p_t \\ &+ \left(1 + \left(1 - \tau_t^k\right) r_t\right) a_{i,t}, \quad i = i_r, \ldots, I. \end{aligned}$$

The alternative interpretation of a mandatory retirement rule is to consider different labor-income tax rates for individuals of ages above and below i_r. In particular, a confiscatory tax on labor income beyond age i_r is equivalent to compulsory retirement. Both formulations yield the same results. However, when we study the optimal policy, we prefer this alternative interpretation since it considers compulsory retirement as just one more distortionary tax that the fiscal authority can optimize over.

In the benchmark economy, a market equilibrium is a sequence of prices and allocations such that consumers maximize utility (equation 1) subject to their corresponding budget constraints (equations (4) and (5)), given the equilibrium prices; firms maximize profits given prices; the government and the Social Security budgets are balanced (equations (2) and (3)); and markets clear and feasibility is satisfied.

PARAMETERIZATION OF THE BENCHMARK ECONOMY

Demographics

We choose one period in the model to be the equivalent of 5 years. Given our choice of period, we assume households live for 12 periods, so that the economically active life of a household starts at age 20, and we assume that households die with certainty at age 80. In the benchmark economy, households retire in period 10 (equivalent to age 65 in years).

Finally, we assume that the mass of households in each period is the same. All these assumptions imply that in the initial steady state the dependency ratio is 0.33 rather than the 0.21 observed nowadays. The reason is that in our simple environment there is no lifetime uncertainty.

Endowments

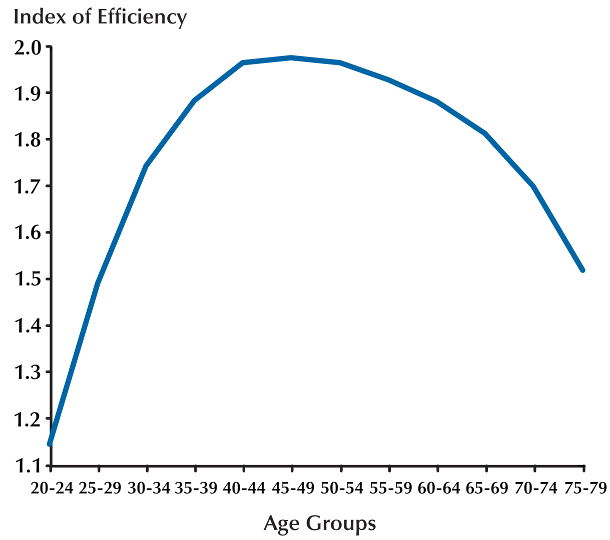
The only endowment that households have is their efficiency units of labor at each period. These are taken from Hansen's (1993) estimates, conveniently extrapolated to the entire lifetime of households (Figure 2).[4]

Government

We assume that in the benchmark economy the government runs two completely independent budgets. One is a Social Security budget that operates on a balanced budget. The payroll tax is taken from the data and is equal to 10.5 percent, which is the Old-Age and Retirement Insurance (OASI).

Figure 2

Age Profile of Efficiency Units of Labor



SOURCE: Hansen (1993).

We exclude a fraction going to disability insurance; the OASDI is 12.4 percent. Our assumptions about the demographics together with the balanced budget condition directly determine the amount of the public retirement pension. It will be 31.5 percent of the average gross labor income.

The level of government consumption is exogenously given. It is financed through a consumption tax, set equal to 5 percent, a marginal tax on capital income equal to 33 percent, and a marginal tax on labor income net of Social Security contributions equal to 16 percent. We have estimated these effective tax rates following Mendoza, Razin, and Tesar (1994). The effective distortion of the consumption-leisure margin is given by $(1-\tau^1)(1-\tau^p)/(1+\tau^c) = 1-0.3$, yielding an effective tax of 30 percent.

The government issues public debt to satisfy its sequential budget constraint.

Calibration: Functional Forms

Households' preferences are assumed to take the form

[4] To avoid sample selection biases, we assume that the rate of decrease of efficiency units of labor after age 65 is the same as in the previous period.

Table 1

Calibration Targets and Parameter Values

Empirical targets	A/Y	IES	Average hours	Contribution of labor income/output	Investment production
Empirical values	3.5	0.5	1/3	0.7	0.12
Parameters	β	σ	γ	α	δ
Calibrated values	1.003	4	0.327	0.3	0.0437

(6) $$\sum_{i=1}^{I} \beta^{i-1} \frac{\left(c_i^{\gamma}\left(1-l_i\right)^{1-\gamma}\right)^{1-\sigma}}{1-\sigma},$$

where $\beta > 0$ represents the discount rate, $\gamma \in (0,1)$ denotes the share of consumption in the utility function, and $\sigma > 0$ governs the concavity of the utility function. The implied intertemporal elasticity of substitution of consumption (IES) is equal to $1/(1-(1-\sigma)\gamma)$.

Technology has constant returns to scale and takes the standard Cobb-Douglas form $Y_t = K_t^{\alpha} L_t^{1-\alpha}$, where α represents the capital-income share.

Calibration: Empirical Targets

We define *aggregate capital* to be the level of fixed sssets in the Bureau of Economic Analysis statistics. Therefore, our calibration target will be a ratio $K/Y = 3$ in yearly terms. Also, computing the ratio of outstanding (federal, state, and local) government debt to gross domestic product (GDP), we get the following ratio $B/Y = 0.5$ in yearly terms. Depreciation is also taken from the data, which as a fraction of GDP is 12 percent. Another calibration target is an average of one-third of the time of households allocated to market activities. We choose a curvature parameter in the utility function consistent with a coefficient of relative risk aversion in consumption of 2 (alternatively a consumption intertemporal elasticity of substitution of 0.5). Government consumption will be fixed at 18.6 percent of output as in the data. Finally, the capital-income share is taken to be equal to 0.3, as measured in Gollin (2002).

Calibration Results

To calibrate our economy, we proceed as follows. First, we fix the curvature parameter in the utility function to be $\sigma = 4$ and the capital share in the production function $\alpha = 0.3$. Then the discount factor $\beta = 1.003$ is chosen to match a wealth-to-output ratio of 3.5,[5] and the consumption share $\gamma = 0.327$ is chosen to match an average of one-third of the time devoted to working in the market economy. The depreciation rate is chosen so that in equilibrium depreciation is 12 percent of output. Notice that $\sigma = 4$ and $\gamma = 0.327$ together imply a consumption intertemporal elasticity of substitution of 0.5 (constant relative risk aversion of 2). Table 1 summarizes the parameters chosen and the empirical targets that are more related to them.

Using the empirical tax rates and ratio of government consumption to GDP, we derive from the government budget constraint an implied equilibrium government debt of 50 percent of output. This figure is consistent with the average figure in the data. Therefore, the capital-to-output ratio is 3 as desired.

Given this parameterization, Social Security annual payments in the benchmark economy amount to 7.35 percent of GDP and the Social Security implicit debt is equal to 128 percent of annual GDP.

THE GOVERNMENT PROBLEM: THE PRIMAL APPROACH

We use the primal approach of optimal taxation as first proposed by Atkinson and Stiglitz (1980). This approach is based on characterizing the set of allocations that the government can

[5] Notice that in a finite-life framework there is no problem with discount factors larger than 1, and, in fact, empirical estimates often take values as large.

implement with the given policy instruments available. A benevolent fiscal authority chooses the optimal tax burden, taking into account the decision rules of all individuals in the economy and the effect of their decisions on market prices.

Therefore, the government problem amounts to maximizing the social welfare function over the set of implementable allocations together with the status quo constraints (that guarentee Pareto improvements).[6] From the optimal allocations, we can decentralize the economy, finding the prices and the tax policy associated with the optimal policy.

A key ingredient is the derivation of the set of implementable allocations. Effectively, it amounts to using the consumer's Euler condition and labor-supply condition to express equilibrium prices as functions of individual allocations and then substituting these prices in the consumer's intertemporal budget constraint. Any allocation satisfying the implementability condition by construction satisfies the household's first-order optimality conditions, with prices and policies appropriately defined from the allocation. See Chari and Kehoe (1999) for a description of this approach.

To illustrate this procedure, we derive the implementability constraint for a newborn individual. Notice that in our case the fiscal authority has to consider retirement pensions as given and that this is going to introduce a difference with Erosa and Gervais (2002), Garriga (1999), or Conesa and Garriga (2008).

We distinguish two cases: first, retirement pensions are considered as regular labor income and are treated as such from a fiscal point of view; and second, retirement pensions are not subject to taxation. Both cases have different tax policy implications.

RETIREMENT PENSIONS AS TAXABLE LABOR INCOME

For clarity of exposition, we suppress the time subscripts. Consider the household maximization problem for a newborn individual facing equilibrium prices and individual specific tax rates on consumption, labor income, and capital income:

$$\max \sum_{i=1}^{I} \beta^{i-1} u(c_i, l_i)$$

$$s.t.\ \left(1+\tau_i^c\right)c_i + a_{i+1} \le \left(1-\tau_i^l\right) w\varepsilon_i l_i + \left(1+\left(1-\tau_i^k\right)r\right)a_i,$$

$$i = 1, \dots, i_r - 1$$

$$\left(1+\tau_i^c\right)c_i + a_{i+1} \le \left(1-\tau_i^l\right)\left(w\varepsilon_i l_i + p\right) + \left(1+\left(1-\tau_i^k\right)r\right)a_i,$$

$$i = i_r, \dots, I$$

$$a_1 = 0,\ a_{I+1} = 0,\ c_i \ge 0, l_i \in (0,1).$$

Notice two important features of this formulation. The first one is that individuals of age i_r and older have a retirement pension, denoted by p, as part of their labor income (and it is taxed at the same rate as regular labor income). Second, on retirement individuals could still supply labor in the market.

Denoting by υ_i the Lagrange multiplier of the corresponding budget constraint, the necessary and sufficient first-order conditions for an interior optimum are given by

$$[c_i] \quad \beta^{i-1} u_{c_i} = \upsilon_i\left(1+\tau_i^c\right), \tag{7}$$

$$[l_i] \quad \beta^{i-1} u_{l_i} = -\upsilon_i\left(1-\tau_i^l\right) w\varepsilon_i, \text{ and} \tag{8}$$

$$[a_{i+1}] \quad \upsilon_i = \upsilon_{i+1}\left[1+\left(1-\tau_i^k\right)r\right], \tag{9}$$

together with the intertemporal budget constraint.

Multiplying these conditions by the corresponding variable we get

$$\beta^{i-1} c_i u_{c_i} = \upsilon_i\left(1+\tau_i^c\right)c_i, \tag{10}$$

$$\beta^{i-1} l_i u_{l_i} = -\upsilon_i\left(1-\tau_i^l\right) w\varepsilon_i l_i, \text{ and} \tag{11}$$

$$\upsilon_i a_{i+1} = \upsilon_{i+1}\left[1+\left(1-\tau_i^k\right)r\right]a_{i+1}. \tag{12}$$

Let $p_i = p$ if $i = i_r, \dots, I$, and zero otherwise. Adding equations (10) and (11) over all i,

[6] Throughout the paper, we assume that the government can commit to its policies. This is an important restriction that affects the results. The analysis of a time-consistent reform goes beyond the scope of this paper.

$$\sum_{i=1}^{I}\beta^{i-1}\left[c_i u_{c_i}+l_i u_{l_i}\right]$$
$$=\sum_{i=1}^{I}\beta^{i-1}\upsilon_i\left[\left(1+\tau_i^c\right)c_i-\left(1-\tau_i^l\right)w\varepsilon_i l_i\right]$$
$$=\sum_{i=1}^{I}\upsilon_i\left(1-\tau_i^l\right)p_i,$$

where the second equality comes from using equation (12) and the budget constraints.

Finally, using equation (8) we get

$$\sum_{i=1}^{I}\beta^{i-1}\left[c_i u_{c_i}+l_i u_{l_i}\right]=-\sum_{i=1}^{I}\beta^{i-1}u_{l_i}\frac{p_i}{w\varepsilon_i}$$

or:

$$(13)\qquad \sum_{i=1}^{I}\beta^{i-1}\left[c_i u_{c_i}+u_{l_i}\left(l_i+\frac{p_i}{w\varepsilon_i}\right)\right]=0,$$

where w denotes the marginal product of labor.

Any feasible allocation of consumption and leisure satisfying equation (13) can be decentralized as the optimal behavior of a consumer facing distortionary taxes. These distortionary taxes can be constructed by using the consumer's optimality conditions for the labor and leisure and for the consumption and savings margins. In particular, given an allocation and its corresponding prices, constructed from the marginal product of labor and capital, we can back out the optimal tax on capital and labor income by using the Euler and labor-supply conditions:

$$(14)\qquad u_{c_i}=\frac{1+\tau_i^c}{1+\tau_{i+1}^c}\beta u_{c_{i+1}}\left[1+\left(1-\tau_i^k\right)r\right],\text{ and}$$

$$(15)\qquad -\frac{u_{l_i}}{u_{c_i}}=\frac{1-\tau_i^l}{1+\tau_i^c}w\varepsilon_i.$$

Notice that in this case the optimal policy is not uniquely determined. Labor and consumption taxation are equivalent in the sense that they determine the same distortionary margin. Also, the taxation of capital income is equivalent to taxing consumption at different times at different rates. In practice, this implies that one of the instruments is redundant. For example, we could set consumption taxes to zero (or to any other constant) and decentralize the allocation using only labor- and capital-income taxes by solving a system of two equations (14) and (15) in two unknowns, τ_i^k and τ_i^l.

Finally, directly using the consumer's budget constraints, we could construct the corresponding sequence of assets. That way we would have constructed an allocation that solves the consumer's maximization problem.

The primal approach of optimal taxation simply requires maximizing a social welfare function over the set of implementable allocations—subject to the feasibility constraint, an implementability condition such as equation (13) for the newborn cohorts, and additional implementability constraints for each cohort alive at the beginning of the reform. We also impose that allocations must provide at least as much utility as in the initial steady state of our economy. The allocation implied by the optimal policy can be decentralized with distortionary taxes in the way we have just outlined.

NONTAXABLE RETIREMENT PENSIONS

If pensions are not taxable, the maximization problem of the households is given by

$$\max\sum_{i=1}^{I}\beta^{i-1}u\left(c_i,l_i\right)$$
$$s.t.\ \left(1+\tau_i^c\right)c_i+a_{i+1}\leq\left(1-\tau_i^l\right)w\varepsilon_i l_i+\left(1+\left(1-\tau_i^k\right)r\right)a_i,$$
$$i=1,\ldots,i_r-1$$
$$\left(1+\tau_i^c\right)c_i+a_{i+1}\leq\left(1-\tau_i^l\right)w\varepsilon_i l_i+p+\left(1+\left(1-\tau_i^k\right)r\right)a_i,$$
$$i=i_r,\ldots,I$$
$$a_1=0,\ a_{I+1}=0,\ c_i\geq 0,\ l_i\in\left(0,1\right).$$

Consequently, through the same procedure used as before we can obtain the expression

$$\sum_{i=1}^{I}\beta^{i-1}\left[c_i u_{c_i}+l_i u_{l_i}\right]$$
$$=\sum_{i=1}^{I}\beta^{i-1}\upsilon_i\left[\left(1+\tau_i^c\right)c_i-\left(1-\tau_i^l\right)w\varepsilon_i l_i\right]$$
$$=\sum_{i=1}^{I}\upsilon_i p_i.$$

Substituting for the Lagrange multiplier, we get

$$\text{(16)}\qquad \sum_{i=1}^{I}\beta^{i-1}\left[u_{c_i}\left(c_i-\frac{p_i}{1+\tau_i^c}\right)+l_i u_{l_i}\right]=0.$$

Notice that in this case the implementability constraint does include a tax term in it, τ_i^c. This did not happen before in expression (13). Hence, it is always possible to choose a particular taxation of consumption such that the implementability constraint is always satisfied. The reason is that now the fiscal authority could tax consumption at a high level but still compensate the consumer through other taxes. In the previous case, this strategy was not available since it was impossible to tax away the retirement pensions and compensate the consumers without introducing additional distortions in the system.

Another way to illustrate this simple intuition is by simply looking at the intertemporal budget constraint of the household:

$$\text{(17)}\qquad \sum_{i=1}^{I}\frac{\left(1+\tau_i^c\right)c_i}{R_i}=\sum_{i=1}^{I}\frac{\left(1-\tau_i^l\right)w\varepsilon_i l_i}{R_i}+\sum_{i=1}^{I}\frac{p_i}{R_i},$$

where $R_1=1,\ R_i=\prod_{s=2}^{i}\left[1+\left(1-\tau_s^k\right)r_s\right]$.

Let $\tau_i^c=\tau^c$. We impose the same taxation of consumption at each point in time of the lifetime of an individual. Then we could rewrite equation (17) as

$$\sum_{i=1}^{I}\frac{c_i}{R_i}=\sum_{i=1}^{I}\frac{1-\tau_i^l}{1+\tau^c}\frac{w\varepsilon_i l_i}{R_i}+\frac{1}{1+\tau^c}\sum_{i=1}^{I}\frac{p_i}{R_i}.$$

Clearly, one could choose any desired level of taxation of τ^c and still introduce no distortion in the consumption-leisure margin by choosing $\tau_i^l=\tau^l=\tau^c$. Effectively, τ^c would act as a lump sum tax.

Therefore, under this new scenario the planner could decentralize a first-best allocation by strategically setting consumption taxes to replicate lump sum taxation.

Notice that this strategy cannot be replicated for the case when retirement pensions are taxable as regular labor income, since the equivalent of equation (17) would be

$$\text{(18)}\qquad \sum_{i=1}^{I}\frac{c_i}{R_i}=\sum_{i=1}^{I}\frac{1-\tau_i^l}{1+\tau^c}\frac{w\varepsilon_i l_i}{R_i}+\sum_{i=1}^{I}\frac{1-\tau_i^l}{1+\tau^c}\frac{p_i}{R_i},$$

and hence the fiscal authority is forced to introduce a distortionary wedge in the consumption-leisure margin when trying to implement lump sum taxation as before.

We are interested in distortionary tax responses to demographic shocks. Consequently, we focus on the scenario where the fiscal treatment of retirement pensions has to be the same as the one of regular labor income. However, we compare the outcomes, in terms of welfare, with the ones that could be obtained if the government could implement lump sum taxation.

THE RAMSEY PROBLEM

We assume that in period $t = 1$ the economy is in a steady state with a PAYG Social Security system and that no demographic shock or government intervention has been anticipated by any of the agents in the economy. The expected utility for each cohort remaining in the benchmark economy is given by

$$\bar{U}_j=\sum_{s=j}^{I}\beta^{s-j}u\left(\hat{c}_s,\ 1-\hat{l}_s\right),$$

where $\hat{c}_s$, $\hat{l}_s$ are steady-state allocations of cohort s. At the beginning of period 2, the demographic shock is known, and then in response to it the optimal policy from then on is announced and implemented. We require that the fiscal authority guarantees to everybody at least the level of utility of the benchmark economy so that the resulting policy reform constitutes a Pareto improvement. This participation constraint will ensure that the optimal response to a demographic shock generates no welfare losses (neither for the initially alive nor the unborn).

Notice that we are imposing a very strong participation constraint, since we require that nobody is worse off relative to a benchmark in which actual fiscal policies would have been sustainable forever (i.e., the initial steady state). Alternatively, we could have postulated different arbitrary policy responses to the demographic shock generating welfare losses for some generations and then improved on those. Clearly, our specification imposes stronger welfare requirements and is independent of any arbitrary nonoptimal policy we might have chosen instead. Besides, the main conclusion in the literature is that no matter what policy you choose, somebody will have to pay the cost of the demographic shock. We show this is not necessarily the case.

The government objective function is a utilitarian welfare function of all future newborn individuals, where the relative weight that the government places on present and future generations is captured by the geometric discount factor $\lambda \in (0,1)$, and $U(c^t, l^t)$ represents the lifetime utility of a generation born in period t.

Conditional on our choice of weights placed on different generations,[7] the Ramsey allocation is the one that solves the following maximization problem:

$$\max \sum_{t=2}^{\infty} \lambda^{t-2} U\left(c^t, l^t\right),$$

$$(19) \qquad s.t. \sum_{i=1}^{I} \mu_{i,t} c_{i,t} + K_{t+1} - (1-\delta) K_t + G_t \le F\left(K_t, \sum_{i=1}^{I} \mu_{i,t} \varepsilon_i l_{i,t}\right) \quad t \ge 2,$$

$$(20) \qquad \sum_{i=1}^{I} \beta^{i-1} \begin{bmatrix} c_{i,t+i-1} u_{c_{i,t+i-1}} \\ + u_{l_{i,t+i-1}} \left(l_{i,t+i-1} + \dfrac{p_i}{F_{2,t+i-1} \varepsilon_i} \right) \end{bmatrix} = 0 \quad t \ge 2,$$

$$(21) \qquad \sum_{s=i}^{I} \beta^{s-i} \left[c_{s,s-i+2} u_{c_{s,s-i+2}} + u_{l_{s,s-i+2}} \left(l_{s,s-i+2} + \frac{p_i}{F_{2,s-i-2} \varepsilon_i} \right) \right] = \frac{u_{c_{i,2}}}{1+\tau_2^c} \left[\left(1 + \left(1 - \tau^k \right) r_2 \right) \bar{a}_{i,2} + p_i \right], \quad i = 2, \ldots, I,$$

$$(22) \qquad \sum_{s=i}^{I} \beta^{s-i} u\left(c_{s,s-i+2},\, 1 - l_{s,s-i+2}\right) \ge \bar{U}_i, \quad i = 2, \ldots, I,$$

and

$$(23) \qquad U\left(c^t, l^t\right) \ge \bar{U}_1, \quad t \ge 2.$$

Constraint (19) is the standard period resource constraint. Constraint (20) is the implementability constraint for each generation born after the reform is implemented and is exactly the one derived in equation (13). This equation reveals that the government faces a trade-off when determining the optimal labor-income tax of the older generations. A higher labor-income tax is an effective lump sum tax on Social Security transfers, but it also reduces the incentives of the older generations to supply labor in the market. The optimal policy will have to balance these opposite forces. Constraint (21) represents the implementability constraints for those generations alive at the beginning of the reform, where τ^k is the benchmark tax on capital income, which is taken as given and $\bar{a}_{i,2}$ are the initial asset holdings of generation i. Notice that taking τ^k as given is not an innocuous assumption, since that way we avoid confiscatory taxation of the initial wealth. Finally, constraints (22) and (23) guarantee that the policy chosen makes everybody at least as well off as in the benchmark economy. In particular, given that the government objective function does not include the initial s generations, equation (22) will be binding.

This formulation imposes some restrictions, since it rules out steady-state golden-rule equilibria. Also, the initial generations alive at the beginning of the reform are not part of the objective function and appear only as a policy constraint.

[7] We are identifying one Pareto-improving reform, but it is not unique. Placing different weights on generations or the initial old would generate a different distribution of welfare gains across agents.

An equivalent formulation would include the initial s generations in the objective function with a specific weight λ_s, where the weight is chosen to guarantee that the status quo conditions for each generation are satisfied.

The policymaker discounts the future at the exponential rate λ. The Pareto-improving nature of the reform implies that the rate λ has to be big enough to satisfy the participation constraints of all future generations. In particular, if λ were too low, then the long-run capital stock would be too low, and future generations would be worse off than in the benchmark economy. That restricts the range of admissible values for λ.

Within a certain range, there is some discretion in the choice of this parameter, implying a different allocation of welfare gains across future generations. To impose some discipline, we choose λ so that the level of debt in the final steady state is equal to that of the benchmark economy and all debt issued along the transition is fully paid back before reaching the new steady state. Our choice of the planner's discount factor, the parameter $\lambda = 0.957$, implies the full repayment of the level of debt issued in response to the demographic shock. That does not mean that the ratio of debt to output will be the same in the final steady state, since output does change.

FURTHER CONSTRAINTS IN THE SET OF TAX INSTRUMENTS

We impose additional restrictions in the set of fiscal instruments available to the fiscal authority. This can be done by using the consumer's first-order conditions in order to rewrite fiscal instruments in terms of allocations and then imposing additional constraints on the Ramsey allocations.

The regime we investigate is one in which capital-income taxes are left unchanged relative to the benchmark. Reformulating this constraint in terms of allocations, we need to impose

$$\begin{aligned}(24)\quad & \frac{u_{c_{1,t}}}{u_{c_{2,t+1}}} = \frac{u_{c_{2,t}}}{u_{c_{3,t+1}}} = \ldots = \frac{u_{c_{I-1,t}}}{u_{c_{I,t+1}}} \\ & = \beta\left[1+\left(1-\tau^{k}\right)\left(f_{k,t+1}-\delta\right)\right], \ \ t \geq 2.\end{aligned}$$

We introduce this constraint since we want to analyze an environment in which the reforms involve only changing the nature of labor-income taxation so that welfare gains are accrued only because of the change in the nature of the financing of retirement pensions rather than a more comprehensive reform involving also changes in the nature of capital-income taxation. Moreover, as Conesa and Garriga (2008) show, the additional welfare gain of reforming capital-income taxation is very small.

With such a constraint, the only instruments available to the fiscal authority will be the taxation of labor income and government debt.

A TRANSITORY DEMOGRAPHIC SHOCK

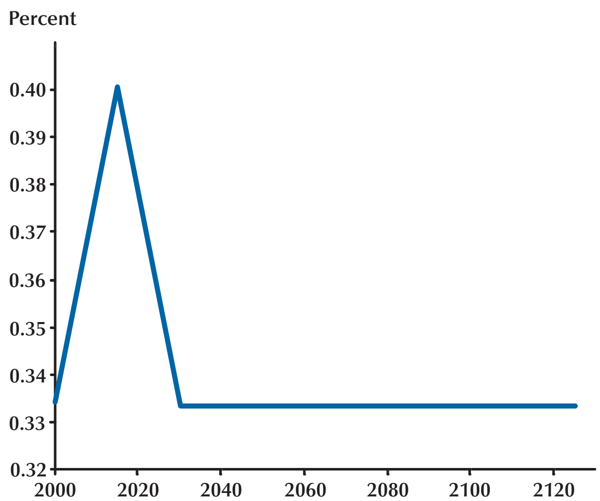
In our experiment, we introduce an unexpected transitory demographic shock, capturing the idea that an increase in the dependency ratio is going to break down the sustainability of the Social Security system we had in the initial steady state of our benchmark economy.

The reason that we want to model it as an unexpected shock is that we want to investigate the optimal response from now on, instead of focusing on what we should have done in advance of an expected shock.

Since introducing realistic demographic projections would imply having to change substantially the demographic structure of our framework, we choose a very simple strategy. We simply increase the measure of retiring individuals for three consecutive periods. Notice that the demographic shock is transitory, in the sense that for three periods (equivalent to 15 years) we face raising dependency ratios, and then for another three periods the dependency ratio falls until reaching its original level and staying there forever. We chose this specification of the demographic shock for computational convenience, since otherwise the model would imply changes in the age structure over time. The alternative would have been an environment where at some point the final age permanently increases reflecting an increase in life expectancy. This raises some com-

Figure 3

Evolution of the Dependency Ratio for Simulated Demographic Shock



putational problems, especially if individuals could forecast the demographic evolution and form expectations about future paths of government action. Hence, the benchmark economy would not be a steady state anymore, and the state of the economy at the benchmark date would be fully driven by arbitrarily chosen expectations. Figure 3 illustrates the evolution of the dependency ratio over time.

We have arbitrarily chosen to label the initial steady state in period 1 as the year 2000, and the demographic shock will be observed and fully predictable at the beginning of period 2 (the year 2005). Hence, the results that follow imply that the policy response from 2005 on is publicly announced and implemented at the beginning of 2005.

Notice that both individuals and the government are assumed to be surprised by the demographic shock. The government learns that given the demographic evolution the system is not sustainable and then implements a policy that rationalizes the financing of pensions. Not only will the government optimally respond to the demographic shock guaranteeing the financial sustainability of pensions in a Pareto-improving way, but moreover the government will permanently change the financing scheme of pensions, hence generating long-run welfare gains relative to the benchmark economy. Our exercise is silent about the reasons why any collective decision process would have resulted in such a distortionary financing scheme in the first place. Indeed, the demographic shock in our exercise triggers the government response, but there is no clear reason why the government should not reform the system in the first place even in the absence of a demographic shock, purely for efficiency considerations. This is exactly what Conesa and Garriga (2008) do in an environment where the government is not constrained to guarantee the pensions promised in the past.

DISCUSSION OF RESULTS

The optimal reform is obtained by solving the maximization problem as stated in the previous section, with the only difference that we have introduced equation (24) as an additional constraint.

We find that the optimal financing scheme implies differential labor-income taxation across age. Why would the government choose to tax discriminate? The critical insight is that when individuals exhibit life-cycle behavior, labor productivity changes with the household's age and the level of wealth also depends on age. As a result the response of consumption, labor, and savings decisions to tax incentives varies with age as well. On the one hand, older cohorts are less likely to substitute consumption for savings as their remaining life span shortens. On the other hand, older households are more likely to respond negatively to an increasing labor-income tax than younger cohorts born with no assets, since the elasticity of labor supply is increasing in wealth. Therefore, the optimal fiscal policy implies that the government finds it optimal to target these differential behavioral elasticities through tax discrimination.

Figure 4 describes the evolution of the average optimal taxes along the reform. We decentralize the resulting allocation leaving consumption taxes unchanged, even though it is possible to decentral-

Figure 4

Evolution of Average Taxes

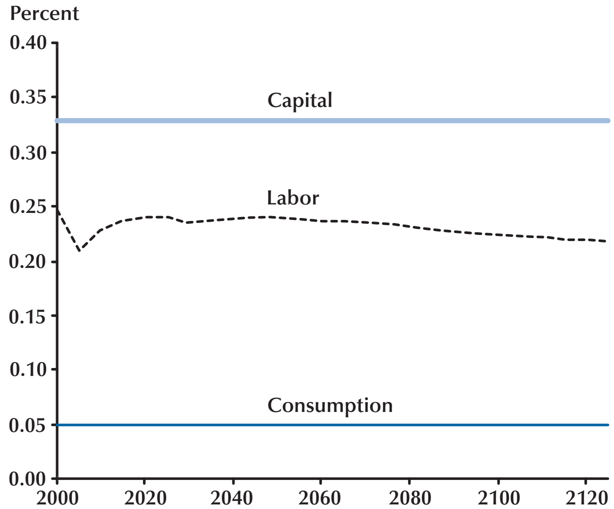
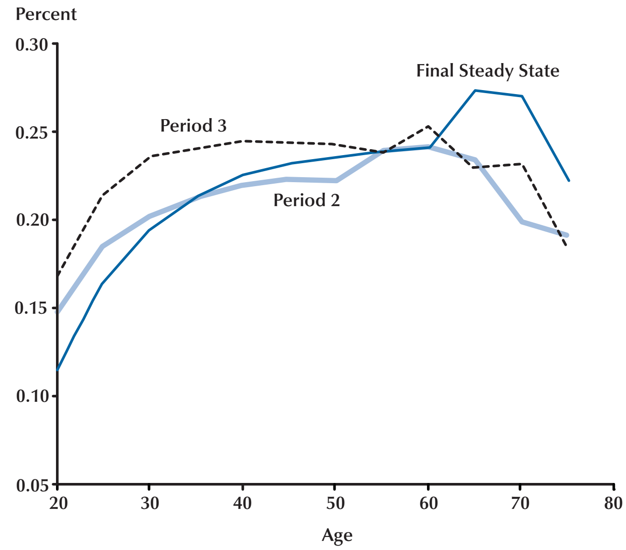


Figure 5

Labor-Income Tax Rates Across Different Cohorts at Different Times



ize the same allocation in alternative ways. In particular, we could set consumption taxes to zero and increase labor-income taxes so that they are consistent with the optimal wedge chosen by the government.

In displaying the results, we arbitrarily label the year 2000 to be the steady state of the benchmark economy, and the reform is announced and implemented the following period (in 2005). Remember that a period in the model is 5 years.

Labor-income taxes are substantially lowered the first period following the reform (the combined impact of labor-income and payroll taxes was a 24.8 percent effective tax on labor in the benchmark), but then they are increased to repay the initial debt issued and reach a new long-run equilibrium around 22 percent on average.

Figure 5 displays its distribution across age at different points in time. The optimal labor-income tax rate varies substantially across cohorts. In the final steady state, the optimal labor-income tax schedule is concave and increasing as a function of age, up to the point at which individuals start receiving a pension. On retirement, the taxation of labor income (remember that retirement pensions are taxed at the same rate as regular labor income) is higher. This feature reflects the tension between the incentives for the fiscal authority to tax away the retirement pensions and the distortions that introduces on labor supply.

Intuitively, the fiscal authority introduces such labor-income tax progressivity to undo the intergenerational redistribution in favor of the older cohorts that the Social Security system is generating.

As a result of this new structure of labor-income taxation, individuals will provide very little labor supply after age 65 and almost none in the last period, as shown in Figure 6. Notice that the shape of labor supply is not dramatically changed with the reform, except for the fact that individuals would still provide some labor while receiving a retirement pension. However, the amount of labor supplied by the oldest cohorts is quite small.

The initial tax cuts, together with the increasing financial needs to finance the retirement pensions, necessarily imply that government debt has to increase in the initial periods following the reform.

Figure 6

Labor Supply Across Different Cohorts at Different Times

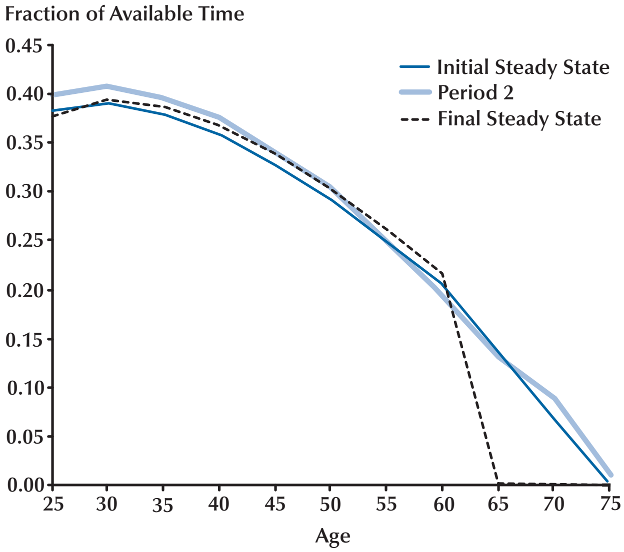


Figure 7

Evolution of Debt-to-GDP Ratio

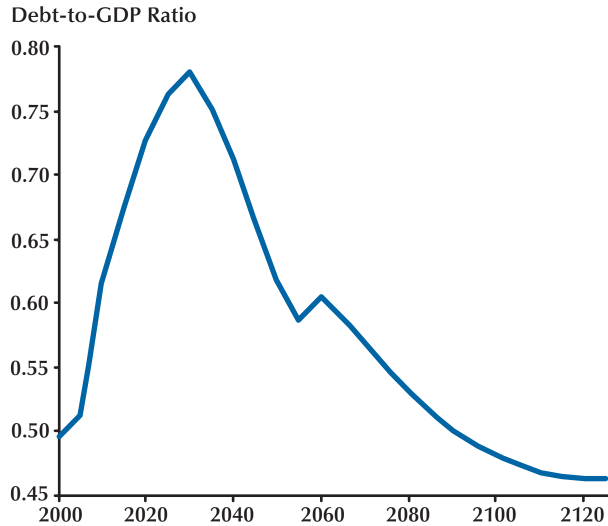
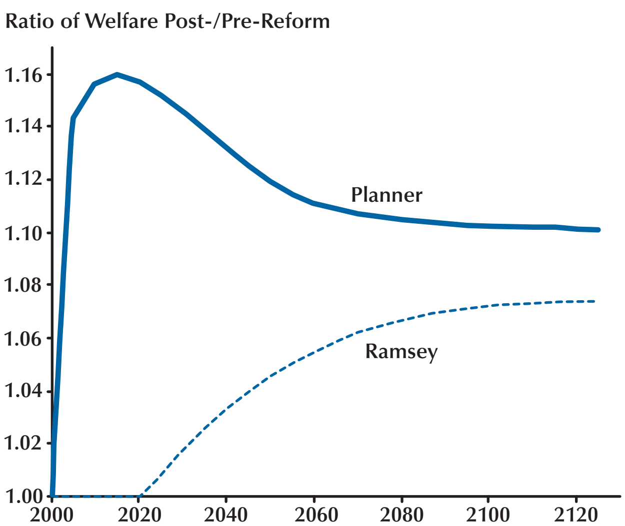


Figure 8

Welfare Gains of Newborn Generations



Next, Figure 7 displays the evolution of government debt over GDP associated with the optimal reform. To finance retirement pensions, government debt would increase up to 77 percent of annual GDP (relative to its initial 50 percent). Later on, this debt will be progressively repaid.

Overall, such a reform generates welfare gains only for those cohorts born once the demographic shock is over. However, the optimal response guarantees that the cohorts initially alive and those born during the shock enjoy the same level of utility as in the benchmark economy. Notice that by construction the initial old were not included in the objective function, and as a consequence the constraint to achieve at least the same utility level as in the benchmark economy has to be necessarily binding. This was not the case for new generations born during the demographic shock since they were included in the objective function of the fiscal authority. Yet the optimal policy response implies that the constraint will be binding, and only after the demographic shock is over will newborn cohorts start enjoying higher welfare. The welfare gains accruing to newborns are plotted in Figure 8.

The optimal response associated with the sustainable policy contrasts with the findings where policies are exogenously specified as in De Nardi, Imrohoroglu, and Sargent (1999), where the initial cohorts are worse off, and Jeske (2003), where the baby boomers and the grandchildren of the baby boomers suffer welfare losses. In our economy, the cost of the shock is distributed over the cohorts initially alive and those generations born during the shock. Remember that the latter do enter the government's objective function, and hence the planner would be happy to allocate some welfare gains to these generations if it were possible.[8]

Notice that the welfare gains associated with the reform just discussed, labeled as "Ramsey" in Figure 8, are much smaller than those associated with the first-best allocation, labeled as "Planner."

By construction, we have prevented the fiscal authority from lump sum taxing the retirement pensions. If we were to allow the fiscal authority to tax differently retirement pensions from regular labor income, the fiscal authority would choose to do so imposing on pensions taxes higher than 100 percent, effectively replicating a system with lump sum taxes. Notice that the welfare gains from doing so (labeled as "Planner") would be much higher, especially for the initial generations. This comparison indicates that the welfare costs of having to use distortionary taxation are very high, especially at the initial periods of the reform.

CONCLUSION

In this paper, we have provided an answer to a very simple and policy-relevant question: What should be the optimal response to an unanticipated transitory demographic shock in Social Security financing? To answer this question, we use optimal fiscal policy to determine the optimal way to finance some promised level of retirement pensions through distortionary taxation. In our experiment, the presence of a demographic shock renders the actual way of financing the Social Security system unsustainable, and our approach endogenously determines how to accommodate this shock, at the same time that the pension financing scheme is permanently changed to reduce distortions.

We find that the government can design a Pareto-improving reform that exhibits sizeable welfare gains in the distant future, after the demographic shock is over. This shows that the pressure induced by the demographic shock is substantial, since the reduction of the existing large distortions only prevents welfare losses but does not generate welfare gains until further away in the future. Our approach explicitly provides quantitative policy prescriptions toward the policy design of future and maybe unavoidable Social Security reforms.

The optimal response consists of the elimination of compulsory retirement, decreasing labor-income taxation of the young, and a temporary increase of government debt to accommodate the higher financial needs generated by the increase in the dependency ratio.

REFERENCES

Atkinson, Anthony B. and Stiglitz, Joseph E. *Lectures in Public Economics*. New York: McGraw-Hill, 1980.

Boldrin, Michele and Montes, Ana. "The Intergenerational State: Education and Pensions." *Review of Economic Studies*, July 2005, *72*(3), pp. 651-64.

Boldrin, Michele and Rustichini, Aldo. "Political Equilibria with Social Security." *Review of Economic Dynamics*, January 2000, *3*(1), pp. 41-78.

Chari, V. V. and Kehoe, Patrick J. "Optimal Fiscal and Monetary Policy," in John B. Taylor and Michael C. Woodford, eds., *Handbook of Macroeconomics*. Volume 1C. Amsterdam: Elsevier, 1999, pp. 1671-745.

Conesa, Juan Carlos and Garriga, Carlos. "Optimal Fiscal Policy in the Design of Social Security Reforms." *International Economic Review*, February 2008, *49*(1), pp. 291-318.

[8] This result shows how large the pressure induced by the demographic shock is. This is especially important since our demographic shock is much less severe than expected even under the most optimistic scenario (compare Figures 1 and 3) and the level of distortions present in our benchmark economy is very high. Hence, our exercise is biased toward generating large welfare gains.

Cooley, Thomas F. and Soares, Jorge. "A Positive Theory of Social Security Based on Reputation." *Journal of Political Economy*, February 1999, *107*(1), pp. 135-60.

De Nardi, Mariachristina; Imrohoroglu, Selahattin and Sargent, Thomas J. "Projected U.S. Demographics and Social Security." *Review of Economic Dynamics*, July 1999, *2*(3), pp. 576-615.

Diamond, P. A. "National Debt in a Neoclassical Growth Model." *American Economic Review*, 1965, *55*(5), pp. 1126-50.

Erosa, Andrés and Gervais, Martin. "Optimal Taxation in Life-Cycle Economies." *Journal of Economic Theory*, August 2002, *105*(2), pp. 338-69.

Escolano, Julio. "Optimal Taxation in Overlapping-Generations Models." Unpublished manuscript, University of Minnesota, 1992.

Feldstein, Martin and Liebman, Jeffrey B. "Social Security," in Alan J. Auerbach and Martin Feldstein, eds., *Handbook of Public Economics*. Volume 4. Amsterdam: Elsevier, pp. 2245-324.

Gale, David. "Pure Exchange Equilibrium of Dynamic Economic Models." *Journal of Economic Theory*, February 1973, *6*(1), pp. 12-36.

Garriga, Carlos. "Optimal Fiscal Policy in Overlapping Generations Models." Unpublished manuscript, University of Barcelona, 1999.

Gollin, Douglas. "Getting Income Shares Right." *Journal of Political Economy*, April 2002, *110*(2), pp. 458-74.

Grossman, Gene M. and Helpman, Elhanan. "Integenerational Redistribution with Short-Lived Governments." *Economic Journal*, September 1998, *108*(450), pp. 1299-329.

Hansen, Gary D. "The Cyclical and Secular Behaviour of the Labour Input: Comparing Efficiency Units and Hours Worked." *Journal of Applied Econometrics*, January-March 1993, *8*(1), pp. 71-80.

Jeske, Karsten. (2003), "Pension Systems and Aggregate Shocks." Federal Reserve Bank of Atlanta Economic *Review*, First Quarter 2003, *88*(1), pp. 15-31; www.frbatlanta.org/filelegacydocs/erq103_jeske.pdf.

Mendoza, Enrique G.; Razin, Assaf and Tesar, Linda L. "Effective Tax Rates in Macroeconomics: Cross-Country Estimates of Tax Rates on Factor Incomes and Consumption." *Journal of Monetary Economics*, December 1994, *34*(3), pp. 297-323.

Federal Reserve Bank of St. Louis
P.O. Box 442
St. Louis, MO 63166-0442