The Journal of Economic Perspectives

A journal of the American Economic Association

Summer 2018

The Journal of
Economic Perspectives

A journal of the American Economic Association

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

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

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, USA.
Printed by LSC Communications, Owensville, Missouri, 65066, USA.
No responsibility for the views expressed by the authors in this journal is assumed by the editors or by the American Economic Association.
THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Summer 2018, Vol. 32, No. 3. The *JEP* is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $20.00, $30.00, or $40.00 depending on income; for an additional $15.00, you can receive this journal in print. E-reader versions are free. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 32 • Number 3 • Summer 2018

Symposia

Macroeconomics a Decade after the Great Recession

Incentives in the Workplace

Articles

Feature

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 3–30

What Happened: Financial Factors in the Great Recession

Mark Gertler and Simon Gilchrist

At least since the Great Depression, major economic calamities have altered the course of research in macroeconomics. The recent global financial crisis is no exception. At the onset of the crisis, the workhorse macroeconomic models assumed frictionless financial markets. These frameworks were thus not able to anticipate the crisis, nor to analyze how the disruption of credit markets changed what initially appeared like a mild downturn into the Great Recession. Since that time, an explosion of both theoretical and empirical research has investigated how the financial crisis emerged and how it was transmitted to the real sector. The goal of this paper is to describe what we have learned from this new research and how it can be used to understand what happened during the Great Recession. In the process, we also present some new empirical work.

This paper is organized into three main parts. We begin with an informal description of the basic theory and concepts, including new developments. This work emphasizes the role of borrower balance sheets in constraining access to credit when capital markets are imperfect. Much of the pre-crisis research focused on constraints facing nonfinancial firms. The events of the Great Recession, however, necessitated shifting more attention to balance sheet constraints facing households

■ *Mark Gertler is Henry and Lucy Moses Professor of Economics, New York University, New York City, New York. Simon Gilchrist is a Professor of Economics, New York University, New York City, New York. Both authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are mark.gertler@nyu.edu and sg40@nyu.edu.*

[†] For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.3

doi=10.1257/jep.32.3.3

and banks. In addition, the crisis brought into sharp relief the need to capture the nonlinear dimension of the financial collapse, prompting a new wave of research.

The next section describes the main events of the financial crisis through the lens of the theory. To tell the story, we also make use of the new wave of empirical research that has sharpened our insights into how the crisis unfolded. In this regard, the literature has been somewhat balkanized with some work focusing on household balance sheets and others emphasizing banks. We argue that a complete description of the Great Recession must take account of the financial distress facing both households and banks and, as the crisis unfolded, nonfinancial firms as well.

We then present some new evidence on the role of the household balance sheet channel versus the disruption of banking. We examine a panel of quarterly state-level data on house prices, mortgage debt, and employment along with a measure of banking distress. Then exploiting both panel data and time series methods, we analyze the contribution of the house price decline, versus the banking distress indicator, to the overall decline in employment during the Great Recession. We confirm a common finding in the literature that the household balance sheet channel is important for regional variation in employment. However, we also find that the disruption in banking was central to the overall employment contraction.

Background Theory and Basic Concepts

In this section, we describe how contemporary macroeconomic models capture the interaction between the financial and real sectors (for recent surveys, see Gertler and Kiyotaki 2011; Brunnermeier, Eisenbach, and Sannikov 2013). Though the models differ in detail, they share several key features: The strength of a borrower's balance sheet, measured by the value of assets net of debt (or "net worth"), affects access to credit and thus the ability to spend. In turn, financial crises are periods where borrower balance sheets contract sharply, leading to a significant disruption of credit flows. Significant declines in spending and economic activity then follow.

Much of the early literature focused on the effect of balance sheet constraints on nonfinancial firms. However, as Bernanke and Gertler (1995) note, the theory applies equally well to households and banks. Indeed, financial distress arose in all three sectors in the recent crisis, as we will elaborate.

The External Finance Premium

The connection between balance sheet strength and credit access arises when frictions impede borrowing and lending. Absent such frictions, a borrower's financial strength is irrelevant to the real investment decision (in an application of the Miller/Modigliani theorem). As a result, with perfect markets the cost of raising funds externally equals the opportunity cost of lending out internal funds.[1]

[1] By external funds, we refer to imperfectly collateralized borrowing. Perfectly collateralized borrowing is effectively the same as using internal funds.

A common way to make financial market frictions endogenous is to introduce an agency problem between borrowers and lenders. There are two basic approaches: either postulating some type of informational asymmetry that leads borrowers to be more informed than creditors, or assuming that it is costly for creditors to enforce certain contractual commitments made by borrowers. In either scenario, borrowers potentially can gain at the expense of lenders by acting dishonestly. Accordingly, rational lenders in this setting will impose constraints on the terms of lending, like credit limits, collateral requirements, and bankruptcy contingencies. Overall, the agency problem makes raising funds externally more expensive than using internal funds, which Bernanke and Gertler (1989) call the "external finance" premium. Indeed, we will argue that an elevated external finance premium is a common feature of financial crises.

Measurement of the external finance premium depends on the details of the agency problem. In many instances, it can be measured as an explicit wedge between borrowing and lending rates due to factors such as costs of evaluating and monitoring borrowers or a "lemons" premium arising when borrowers are likely better informed about their creditworthiness than are lenders. In other cases, where there is nonprice rationing due to some form of credit limit, covenant restriction, or collateral requirement, the external finance premium is measured as the difference between the "shadow borrowing rate" and the lending rate, where the shadow borrowing rate is the borrower's marginal return to investing. In either case, the external finance premium adds to the cost of capital.[2]

Key to the behavior of the external finance premium is the borrower's balance sheet. In a situation with agency problems, a stronger balance sheet enables the borrower either to self-finance a greater fraction of an investment or to provide more collateral to guarantee the debt. This basic prediction—that credit access improves with the strength of the balance sheet—is characteristic of many real-world financial arrangements, including restrictions that borrowers post down payments, post collateral, and meet certain financial ratios. In any of these cases, a borrower who is able to take a larger stake in the outcome of the investment will have a reduced level of agency conflict with the lender. The external finance premium declines as a consequence.

The Financial Accelerator/Credit Cycle Mechanism and Crises

The link between borrower balance sheets and the external finance premium leads to mutual feedback between the financial sector and real activity. A weakening

[2] It might seem that an alternative approach is to examine the behavior of credit aggregates and then consider the forecasting power of these aggregates for real activity. However, this approach cannot disentangle whether demand or supply is driving the movement in these quantities. Loan demand is likely to vary positively with real activity, leading to a positive correlation between credit quantities and output. Thus, procyclical variation in credit aggregates can arise even when financial market frictions are absent. We do not mean to suggest that the behavior of credit aggregates is uninformative about financial conditions. They can reveal the risk exposure of different sectors, as measured by the degree of leverage. But a measure of the quantity of credit alone does not tell us how tight or loose financial constraints are.

of balance sheets raises the external finance premium, reducing borrowing, spending, and real activity. The decline in real activity reduces cash flows and asset prices, which weakens borrower balance sheets, and so on. This kind of adverse feedback loop was captured originally by the financial accelerator model in Bernanke and Gertler (1989) and Bernanke, Gertler, and Gilchrist (1999) and the credit cycle model in Kiyotaki and Moore (1997).[3] Many contemporary models of financial crises have evolved from this approach.

With a sufficient deterioration of balance sheets, a full-blown financial crisis emerges as external finance premia rise to the point where borrowers are induced to curtail spending sharply. In fact, this combination of weak balance sheets and high external finance premia is characteristic of major financial crises. A rough proxy for the external finance premium is the interest rate spread between the return on a private debt instrument, such as a corporate bond, a mortgage, or commercial paper, and a similar maturity government bond. These spreads tend to widen across the board during crises and did so dramatically during the recent crisis.

This earlier literature focused largely on constraints faced by nonfinancial firms. In the recent crisis, however, it was mainly highly leveraged households and highly leveraged banks that were initially vulnerable to financial distress. Thus, motivated by the seminal empirical work of Mian and Sufi (2014) and Mian, Rao, and Sufi (2013), studies like Eggertsson and Krugman (2012), Justiniano, Primiceri, and Tambalotti (2010), and Guerreri and Lorenzoni (2017) incorporated balance sheet constraints on households. The distress in financial markets induced other studies like Gertler and Kiyotaki (2011), He and Krishnamurthy (2013), and Brunnermeier and Sannikov (2014) to incorporate balance sheet constraints on banks. In these studies, the financial accelerator mechanism remains operative, but the transmission of the crisis through the different sectors of the economy is much closer to what actually occurred. [4]

The Role of Leverage

The exposure of the economy to a financial crisis is closely related to the degree to which borrowers rely on debt. The higher the fraction of financing that is debt, as opposed to equity, the more sensitive the balance sheet becomes to fluctuations in asset prices. For example, consider a borrower that self-finances an asset versus one who self-finances 10 percent and issues debt to finance the rest. A 10 percent

[3] Bernanke's (1983) classic analysis of the role of financial factors in the Great Depression provided motivation for this direction.

[4] Readers interested in some additional examples of macro modeling of financial crises might also look at Geanakopolos (2010), Jermann and Quadrini (2012), Christiano, Motto, and Rostagno (2014), Arellano, Bai, and Kehoe (2016), and Iacoviella (2005). Also, while the modern literature has formalized this theory of financial crises, some of the ideas have an earlier pedigree. For example, Irving Fisher's (1933) debt-deflation theory of the Great Depression held that the weakening of borrower balance sheets stemming from the sharp price deflation during the early 1930s was a significant factor driving the depth and duration of the Depression. The deflation weakened balance sheets because most debts were in nominal terms.

decline in the asset values will leave the former with a 10 percent reduction in net worth, while the latter will be completely wiped out.

The lead-up to the Great Recession saw an unprecedented rise in leverage in both the household and banking sectors. Household leverage was largely in the form of mortgage debt, occurring in the context of a dramatic boom in housing prices. Both investment banks and commercial banks financed the increase in mortgage holdings by mostly short-term debt of their own. The fact that the bank debt was mostly short term also made the system vulnerable to runs, as we discuss shortly. By 2006, the financial positions of both households and banks were highly vulnerable to the decline in house prices that would soon follow.

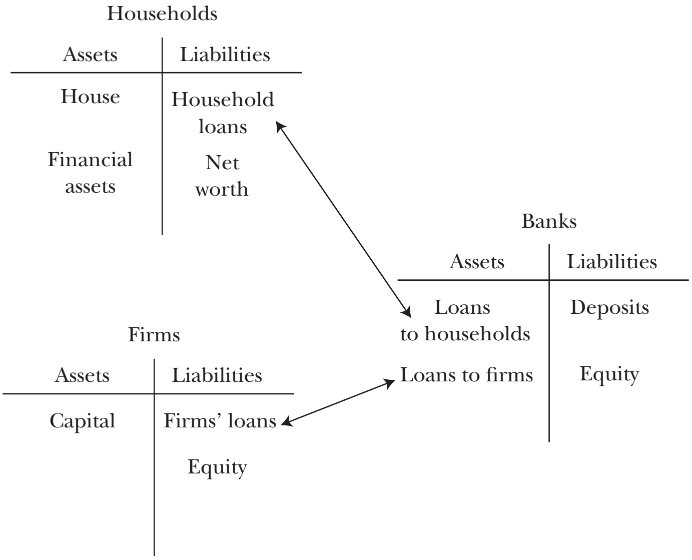
Nonlinear Effects of Financial Crises

Financial crises are highly nonlinear events (for example, see Krishnamurthy, Nagel, and Orlov 2014). Such crises typically feature sharp increases in credit spreads and sharp contractions in asset prices and output. However, booms do not experience a symmetric countermovement of these variables. Further, the sharp contraction of the economy during a financial crisis often occurs without any immediate large nonfinancial shock to the economy, as was the case for the US economy in the last few months of 2008.

The earlier generation of financial accelerator models (Bernanke, Gertler, and Gilchrist 1999) considered loglinear approximations around a deterministic steady state and thus could not capture nonlinear dynamics. Recent literature has addressed the issue in a variety of ways. For example, Mendoza (2010) and He and Krishnamurthy (2014) introduce nonlinearity by allowing balance sheet constraints that bind only during recessions, not booms. To put it another way, the economy during a boom behaves to a large extent as if it had frictionless financial markets. However, a negative disturbance can move the economy into a region where the constraints are binding, amplifying the effect of the shock on the downturn. In a related approach, Brunnermeier and Sannikov (2014) develop a framework where, for precautionary reasons, borrowers reduce spending by more in response to a contraction in the balance sheet than they increase it in response to a strengthening of similar magnitude. These kind of asymmetries can help account for why, during the recent recession, household consumption responded more strongly to contractions in house prices that weakened household balance sheets than to the earlier run-up in housing prices.

More recently, Gertler, Kiyotaki, and Prestipino (2017) develop a framework with bank runs as the key source of nonlinearity. The key element here is whether financial institutions like investment banks are able to roll over their short-term loans. Within this model, in normal times where banks have healthy balance sheets, lenders are confident that even if other creditors do not roll over, the bank has the resources to honor its debt. However, in downturns where bank balance sheets have weakened, lenders can no longer be certain their deposits are safe if other creditors were to withdraw. As a consequence, a self-fulfilling roll-over panic becomes possible, which generates a highly nonlinear rise in credit spreads and contraction in asset prices and output.

Figure 1
Sectoral Balance Sheets



Interdependence of Household, Firm, and Bank Balance Sheets

In analyzing the dynamics of a financial crisis, it is critical to account for the interdependence of balance sheets across sectors. Figure 1 illustrates the interconnection between household, firm, and bank balance sheets. (We simplify for expositional purposes.) For households, assets consist of housing and financial assets. Liabilities are loans from banks, and net worth. Bank assets are loans to households and loans to firms. Bank liabilities are deposits and equity. In turn, loans along with equity are on the liability side of firm balance sheets, while assets consist of capital.

Clearly, the balance sheet position of one sector of the economy will also affect others. Household debt—and mortgage debt in particular—typically surges prior to a financial crisis (for example, Mian, Sufi, and Verner 2017; Shularick and Taylor 2012). The origins of the Great Recession similarly involved a surge in mortgage lending and a boom in house prices and housing construction. As the house price boom began to reverse, household balance sheets weakened, and consumption growth fell.

But mortgages also appear on the asset side of bank balance sheets. Indeed, the lion's share of the growth in mortgages since the late 1990s was created by securitized mortgage loans, which were absorbed by a huge expansion of the thinly capitalized and lighted regulated shadow banking sector. When banks (broadly defined) are subject to financial distress, the flow of credit is impeded to the broad spectrum of nonfinancial borrowers, including firms as well as households.

The Relevance of Constraints on Monetary Policy

The severity of a financial crisis depends critically on the behavior of monetary policy. When the central bank is free to respond, it can (at least partially) offset the effect of the crisis on the cost of credit by reducing interest rates. Conversely, when the hands of the central bank are tied, the crisis is much more likely to spin out of control. The evidence is consistent with this insight. For example, for emerging market economies in the post–World War II period, full-blown financial crises were more likely to occur in countries operating under fixed exchange rates, where monetary policy was not free to adjust, as opposed to countries operating under flexible rates (Kaminsky and Reinhart 1999). Similarly, Eichengreen (1992) and others have shown that during the Great Depression era, countries that freed up their monetary policy by abandoning the gold standard early in the crisis experienced much milder downturns than those that delayed.

For the recent financial crisis, the relevant constraint on monetary policy was the zero lower bound on the nominal interest rate. As financial conditions deteriorated and the economy began contracting in fall 2008, the Federal Reserve quickly reduced short-term interest rates, which reached zero, effectively, by December 2008. From that point on, the Fed's conventional tool was no longer available. The zero lower bound also constrained the other major central banks, including the European Central Bank and the Bank of England. Of course, the Bank of Japan had a much longer experience with the zero lower bound going back to the 1990s.

All of these central banks, led by the Federal Reserve, introduced a variety of unconventional monetary policies to circumvent the constraints of the zero lower bound. The most visible of these policies was large-scale asset purchases ("quantitative easing"), which the Fed introduced after the peak of the crisis in early 2009. This paper is not the place to go into detail about these policies: for a formal analysis of how unconventional monetary policy affects the economy, see Gertler and Karadi (2011) and Curdia and Woodford (2011). However, these unconventional monetary policy interventions are widely credited for helping mitigate the severity of the financial crisis.

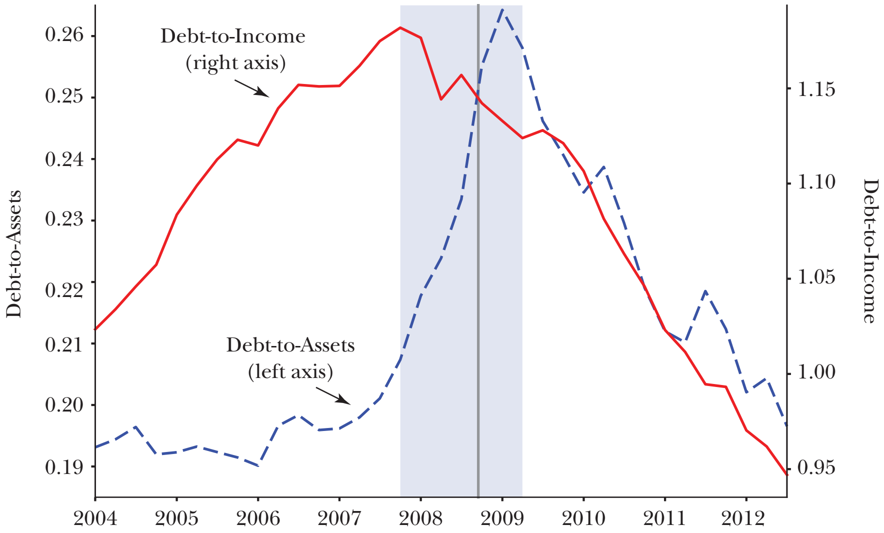
The Financial Crisis through the Lens of the Theory

In this section, we use the theory outlined in the previous section as an organizing framework to identify the role of financial factors in the unfolding of the Great Recession. In particular, we identify how and when balance sheet constraints in each of the three sectors—households, banks, and firms—become relevant. For much of the background material, we rely on Bernanke (2010, 2015), Gorton (2010), Adrian and Shin (2010), and Gertler, Kiyotaki, and Prestipino (2016).

Buildup of Vulnerabilities

The prelude to the financial crisis was an extraordinary housing boom, featuring a dramatic run-up in house prices, residential construction, and mortgage debt. A variety of factors triggered the boom, including a secular decline in long-term interest rates, a relaxation of lending standards, and widespread optimism about future increases in house prices. In addition, increased securitization of mortgages permitted greater separation of the origination function of mortgage

Figure 2
Debt/Income and Debt/Assets: Households



Note: Figure 2 provides information on the household balance sheet over the ten-year period from 2004 through 2014. The shaded area is the time from peak to trough of the Great Recession and the vertical line marks the Lehman Brothers bankruptcy, which is generally considered the epicenter of the financial collapse. The figure portrays two measures of household leverage: the ratio of household debt-to-income (the solid line) and the ratio of household debt-to-assets (the dashed line), where the latter includes the market values of housing and financial wealth.

lending from the funding role. Lightly regulated shadow banks began to displace commercial banks as the primary funders of mortgage-related securities.[5] One example is the rise of asset-backed commercial paper conduits, which held securitized assets such as mortgages and car loans and funded these assets by issuing short-term (for example, 30-day) commercial paper. The cost of mortgage finance declined because these shadow banks did not face the same capital requirements or regulatory oversight as commercial banks.

The housing boom made both households and banks financially vulnerable. Figure 2 provides information on the household balance sheet over the ten-year period from 2004 through 2014. The shaded area is the time from peak to trough of the Great Recession and the vertical line marks the Lehman Brothers bankruptcy, which is generally considered the epicenter of the financial collapse. The figure portrays two measures of household leverage: the ratio of household debt-to-income (the solid line) and the ratio of household debt-to-assets (the dashed line), where the latter includes the market values of housing and financial wealth. From 2004Q1 to the start of the recession, household debt-to-income increased roughly

[5] Government-sponsored enterprises (GSEs) also continued to play a significant role in mortgage lending.

16 percent, fueled mainly by the rapid increase in mortgage debt. Household asset values increased at roughly the same pace as the increase in mortgage debt mainly due to the rapid increase in house prices. The net effect is that the debt-to-assets ratio rose comparatively little until the start of the Great Recession.

By the end of 2007, households were vulnerable to the sharp decline in asset values that would follow. Housing prices peaked at the end of 2006 and then declined more than 25 percent. As a result, the aggregate household leverage ratio—measured by the ratio of debt to assets—increased roughly 25 percent from early 2007 to the business cycle trough. Later in the recession toward the end of 2008, the decline in stock prices also contributed further to the rise in the household leverage ratio. Of course, certain states like California and Florida experienced much sharper declines in house prices and increases in household leverage than the national average.

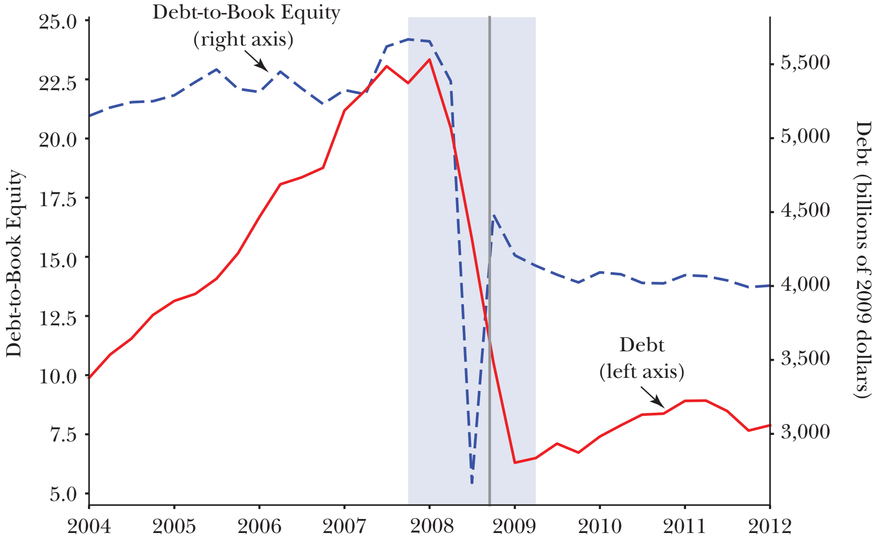
The deterioration of household balance sheets provided a channel through which declining house prices affected household spending and, in turn, economic activity. The weakening of the household balance sheet reduces access to credit, like home equity loans.[6] A substantial literature initiated by the seminal work of Mian, Rao, and Sufi (2013) and Mian and Sufi (2014) has examined the role of the household balance sheet channel during the Great Recession. To identify the strength of this channel, this work exploits the regional variation in house prices and household balance sheets that we alluded to earlier. We return to this issue of estimating effects using regional variation later.

As vulnerabilities in household balance sheets materialized, corresponding vulnerabilities in bank balance sheets emerged as well. Shadow banks grew from intermediating less than 15 percent of credit in the early 1980s to roughly 40 percent on the eve of the Great Recession, an amount on par with commercial banks (see Gertler, Kiyotaki, and Prestipino 2017). Turning to Figure 3, the solid line shows from 2004 to the start of the Great Recession investment banks (a major component of the shadow banking sector) increased their real debt levels by more than 50 percent, mostly as a consequence of financing the rapid expansion in securitized assets by borrowing in short-term credit markets. Because these firms did not face the regulatory capital requirements of traditional banks and because they generally received high marks from the credit ratings agencies like Standard & Poor's, Moody's, and Fitch on the mortgage-related securities that they held, the investment banks tended to operate with much higher leverage ratios than did the commercial banks. Prior to the Lehman Brothers collapse in September 2008, investment banks operated at ratios of debt-to-assets of between 20 and 25, roughly three times the level of commercial banks. Other types of shadow banks, including asset-backed commercial paper issuers and finance companies, similarly operated with high leverage.

[6] The argument in the text requires imperfect financial markets. With perfect financial markets and the ability to borrow freely based on lifetime income, a drop in house prices does not induce a wealth effect on household spending because the decline in house prices is offset by the decline in the cost of housing (assuming that the household continues to reside in the same neighborhood where house prices have declined).

Figure 3

Debt and Debt/Equity: Investment Banks



Note: Figure 3 provides information about the balance sheet behavior of publicly traded investment banks, a major component of the shadow banking sector. The shaded area is the time from peak to trough of the Great Recession and the vertical line marks the Lehman Brothers bankruptcy, which is generally considered the epicenter of the financial collapse.

The increase in the quantity of mortgage debt was accompanied by a decline in the quality. As Bernanke (2015) notes, the riskiest mortgages were issued in 2005 and 2006, at the height of the house price boom. Mortgages that were clearly labeled as risky from the start included both "sub-prime" (issued primarily to low-income borrowers) and also "Alt A" (issued to speculators and/or households taking out second mortgages). In 2005 and 2006, the share of newly issued mortgages that could be classified a priori as risky rose to roughly 40 percent, up from 10 percent in 2002. A general relaxation of lending standards helped to fuel the increase. Also complicating matters is that roughly 30 percent of newly issued mortgages were issued at variable interest rates rate at a time when the Federal Reserve was in the midst of a tightening cycle, adding to their overall risk.

The Unraveling

A combination of declining house prices and increasing short-term interest rates led to an uptick in mortgage defaults in 2007, particularly on low-grade variable rate mortgages issued in 2005 and 2006. In July 2007, the investment bank Bear Stearns defaulted on two of its mutual funds that were exposed to mortgage risk. In August 2007, in the event largely considered to mark the beginning of the crisis, the French bank BNP Paribas suspended withdrawals from funds that also had mortgage exposure risk.

Concern spread quickly about other financial institutions with mortgage risk exposure, particularly those relying heavily on short-term funding. The asset-backed commercial paper market was an early target (as discussed in Kacperczyk and Schnabl 2010; Covitz, Liang, and Suarez 2013). Again, intermediaries in this market funded securitized assets, including pools of mortgages, auto loans and credit card debt, and so on. They funded these assets by issuing short-term commercial paper, using the assets as collateral. Concern about the quality of these assets, however, especially those with mortgage exposure, led suppliers of commercial paper (like money market funds) to either tighten the terms of credit or withdraw from the market completely. The value of asset-backed commercial paper outstanding fell from a peak of $1.2 trillion in June 2007 to $800 billion by the following December.

The way in which the contraction of the asset-backed commercial paper market transmitted to the real economy can be described in terms of the theory presented in the previous section. The reduction in the perceived collateral value of the securities held by asset-backed commercial paper issuers weakened their balance sheets and raised the cost of access to the commercial paper market. Interest rates on asset-backed commercial paper increased relative to Treasury bill rates of similar maturity. Other terms of lending, such as collateral requirements, tightened as well. The increase in funding costs faced by issuers of asset-backed commercial paper in turn raised the cost of credit for mortgages, auto loans, and other types of borrowing that made use of securitized lending.

The collapse of the asset-backed commercial paper market led to the first significant spillover of financial distress to the real sector, contributing to the slowdown in residential investment, automobile demand, and other types of spending that relied on this funding. Benmelech, Meisenzahl, and Ramacharan (2017), for example, present evidence that tightening of credit conditions in the asset-backed commercial paper market accounted for roughly one-third of the overall decline in automobile spending during the crisis.

At the same time, the decline in house prices was weakening household balance sheets, placing downward pressure on consumer spending. In addition, the end of the housing boom meant a sharp drop in residential investment. These factors, along with the disruption of short-term credit markets like asset-backed commercial paper, were sufficient to move the US economy into recession at the end of 2007.

The Federal Reserve responded aggressively to the onset of the recession. It reduced the federal funds interest rate and undertook a variety of measures designed to improve the availability of short-term credit. These measures included making it easier for commercial banks to obtain discount window credit and also making this credit available to investment banks (which had previously been unable to borrow in this way). The Federal Reserve also exchanged government bonds for highly rated private securities to boost the supply of (perfectly) safe assets that could be used to collateralize short-term borrowing. The most dramatic intervention involved the steps taken in the spring of 2008 to prevent solvency problems with Bear Stearns from further disrupting credit markets: The central bank provided funding for JP Morgan's acquisition of Bear Stearns using some of the latter's assets as collateral.

Collapse of the Financial and Real Sectors

Through the summer of 2008, the US economy continued to slow. However, the common perception at the time was that it would experience a downturn similar to the relatively moderate recession of 1990–91, which also featured a banking crisis, though one that involved commercial real estate and commercial banks rather than residential real estate and shadow banks.

In September 2008, however, the second and larger wave of financial distress hit. Lehman Brothers, a much larger investment bank than Bear Stearns, was similarly exposed to mortgage-related risk. A significant decline in the value of its securities holdings weakened its balance sheet and raised the risk to its short-term creditors, from whom it was obtaining virtually all its funding. The Reserve Primary Fund, a large money market mutual fund that held commercial paper issued by Lehman, experienced a run that forced it into liquidation. Runs on other money market funds were only averted when the US Treasury extended deposit insurance to these institutions.

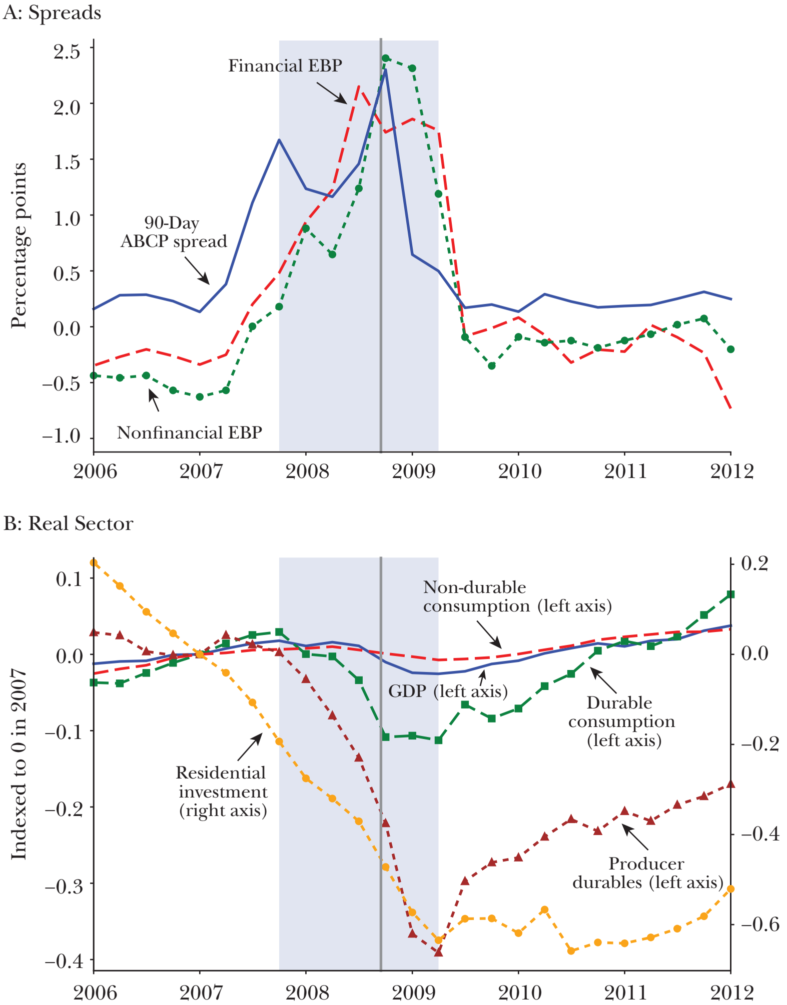
The distress then spread to Lehman's main source of short-term funding, the repo market in which borrowers obtained overnight loans using securities as collateral. The uncertainty about the value of these securities, particularly if there was a hint of mortgage risk exposure, made creditors less willing to accept them as collateral, leading many to pull out of the repo market (for discussion, see Krishnamurthy, Nagel, and Orlov 2014). What emerged were bank runs in the spirit of Diamond and Dybvig (1983), though in markets for wholesale funding (interbank) as opposed to retail funding. In addition, weakening of their balance sheets exposed these institutions to runs, which took the form of a collective failure of creditors to roll over their loans (as in Gertler and Kiyotaki 2015; Gertler, Kiyotaki, and Prestipino 2017).

The Federal Reserve was unable to act as a "lender of last resort" to Lehman because the bank could not offer sufficient collateral. The lack of short-term credit forced Lehman into default. Fearing similar vulnerability, the other major investment banks quickly merged with commercial banks in order to get the regulatory protection afforded to the latter. The contraction in investment banking impeded credit flows, placing further downward pressure on economic activity.

The financial crisis spread like a cancer from the shadow banking sector, which funded mainly securitized assets, to the commercial banking sector. When commercial banks merged with investment banks, they also absorbed a share of the assets funded by the investment banks. But commercial banks were limited in the amount they could absorb by their equity capital in conjunction with capital requirements that limited their leverage ratios well below the level at which the investment banks had operated. An additional source of pressure on commercial banks was losses on securitized assets that they had initiated and sold. Even though the banks sold these assets, they had an implicit commitment to absorb the losses. The losses on mortgage-related assets in turn weakened the balance sheets of commercial banks, disrupting the flow of credit through these institutions. Now bank-dependent borrowers, including many nonfinancial firms and households, also faced increasing credit costs.

Figure 4

Credit Spreads and Economic Activity



Note: Figure 4A portrays the behavior of three key credit spreads: the 90-day asset-based capital spread (ABCP is Asset-Backed Commercial Paper); the Gilchrist and Zakrajsek (2012) excess bond premium (EBP) for nonfinancial companies; and the excess bond premium (EBP) for financial companies. In each case, the spread measures the difference between the return on the security and the return on a government bond of similar maturity. Figure 4B shows the accompanying behavior of the real sector, including GDP and four key components: residential investment, consumer durables, producer durables and nondurable consumption. (All variables are in logs.) The shaded area is the time from peak to trough of the Great Recession and the vertical line marks the Lehman Brothers bankruptcy.

The major disruption of financial intermediation following these events in September 2008 led to a sharp across-the-board contraction in economic activity. Figure 4 illustrates. The top panel portrays the behavior of three key credit spreads:

the 90-day asset-based commercial paper spread; the Gilchrist and Zakrajšek (2012) excess bond premium for nonfinancial companies; and the excess bond premium for financial companies. In each case, the spread measures the difference between the return on the security and the return on a government bond of similar maturity.[7] The spread for asset-backed commercial paper increases roughly 150 basis points from early 2007 to the end of that year, reflecting the problems in that market that developed prior to the onset of the recession. After a slight dip, the asset-backed commercial paper spread increased another 100 basis points in response to the turmoil in the commercial paper market following the Lehman collapse in September 2008. As the turbulence spread to both investment banks and commercial banks, the excess bond premium for financial companies increased to more than 150 basis points in the wake of the Lehman collapse. Finally, the contraction of the shadow banking sector along with the subsequent disruption of commercial banking steadily pushed up credit costs faced by nonfinancial borrowers. As an example, the excess bond premium increased by 275 basis points at the time of the Lehman default.[8]

The bottom panel in Figure 4 shows the accompanying behavior of the real sector, including GDP and four key components: residential investment, consumer durables, producer durables, and nondurable consumption. (All variables are in logs.) The growth rate of GDP moves slightly negative in the early stages of the recession starting in late 2007. Contributing to the initial slowdown is a sharp decline in residential investment as pessimism about future housing prices begins to grow. Financial factors also play a role. Problems in the asset-backed commercial paper market led to upward pressure on the cost of mortgage credit. In addition, as Gilchrist, Siemer, and Zakrajšek (2017) emphasize, the disruption of credit markets also increased borrowing costs for construction companies that were building homes on speculation.

Also contributing to the initial slowdown was a drop in consumer durable demand at the beginning of the recession, largely due to a sharp decline in automobile demand. Here, forces operated through both household and bank balance sheets. Using cross-regional evidence, Mian, Rao, and Sufi (2013) show that the weakening of household balance sheets due to the decline in house prices induced a significant drop in automobile demand. On the other side of the ledger, as we mentioned earlier, Benmelech, Meisenzahl, and Ramacharan (2017) showed that

[7] The excess bond premium is the difference between the yield on an index of nonfinancial corporate bonds and a similar maturity government bond, where the latter is adjusted to eliminate default risk. The idea is to have a pure measure of the excess return that is not confounded by expectations of default. The excess bond premium in the financial sector is constructed in an analogous manner for publicly traded companies in the financial sector.

[8] As emphasized by Adrian, Colla, and Shin (2012) and Becker and Ivashina (2014), the deterioration in the financial health of commercial banks induced many nonfinancial borrowers to switch from bank to public debt markets to obtain credit, placing upward pressure on the excess bond premium. For an early theoretical description of this bank loan supply effect on corporate bond rates, see Kashyap, Stein, and Wilcox (1993).

the disruption of the asset-backed commercial paper market had a significant negative effect on the demand for cars.

Following the Lehman bankruptcy at the end of the third quarter of 2008, the recession turned from mild to major. GDP began a sharp contraction that lasted until the spring of 2009. As credit costs rose across the board, demand fell for both consumer and producer durable goods. Consumer durables dropped roughly 15 percent while producer durables dropped a whopping 35 percent. Financial factors also contributed significantly to the contraction in producer durables. Entering the recession, nonfinancial firms were not directly financially vulnerable to the fall in home prices in the same way that households and (shadow) banks were. They did not (on average) run up their leverage ratios, nor were they directly exposed to house price risk. On the other hand, as the crisis unfolded, equity values dropped significantly, weakening firm balance sheets. Also, the increased strain on commercial banks made access to credit more difficult for nonfinancial firms, as just mentioned.

Figure 5 illustrates how financial distress hit the nonfinancial business sector. Figure 5A plots the debt/equity ratio of the nonfinancial corporate business sector alongside a measure of the external finance premium for nonfinancial companies, specifically the Gilchrist/Zakrasjek excess bond premium we used in Figure 4. Consistent with the theory we described earlier, a higher credit spread is associated with a high leverage ratio.

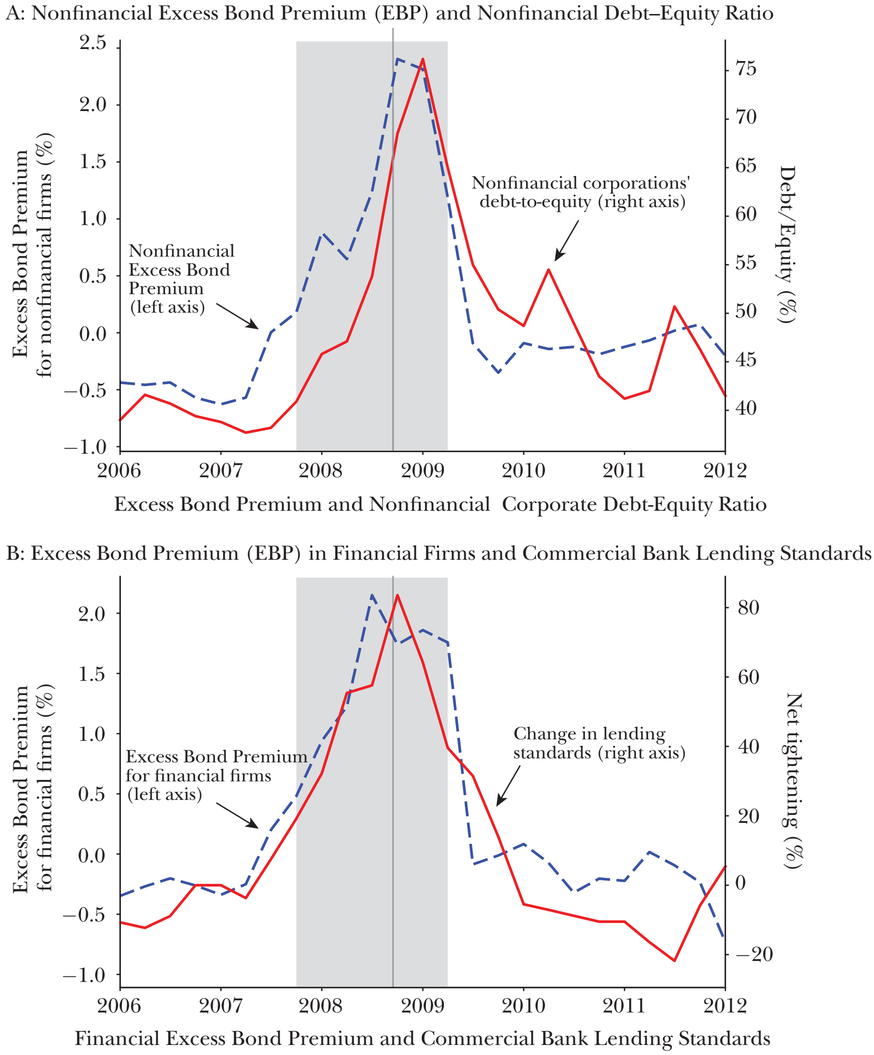
Figure 5B shows how distress in banking may have affected the flow of credit to the nonfinancial business sector. It plots the excess bond premium for financial companies (a measure of the distress facing financial institutions) against an indicator of the tightness of bank credit, which comes from a survey of lending senior loan officers about lending terms. As the figure shows, they are closely correlated. Note also that during the Great Recession, the unusually high degree of tightening in lending standards shown in the survey data is also correlated with the sharp increase in the financial excess bond premium, consistent with the latter being a contributing factor to the former.

Formal panel-data studies also identify a role for financial factors influencing nonfinancial firm behavior. For example, Giroud and Mueller (2015) show that firms that had built up their leverage prior to the Great Recession accounted mainly for the subsequent contraction in employment across regions. As noted earlier, Chodorow-Reich (2014) and Chodorow-Reich and Falato (2017) document that bank health affected the flow of credit to nonfinancial firms. Finally, Gilchrist, Shoenle, Sim, and Zakrajšek (2017) show that liquidity constraints induced a fraction of firms to raise their price markups in order to generate increased cash flow over the near term (at the likely cost of reducing future market share).

The financial and economic contraction following the Lehman bankruptcy in September 2008 induced a massive policy response, including steps aimed at addressing the problems of financial sector intermediation and bank balance sheets. The Federal Reserve quickly reduced the short-term interest rate to zero, but it also pursued a variety of other interventions. Among the most visible was massive purchases of agency mortgage-backed securities financed mainly by issuing

Figure 5

How Distress in Banking May Have Affected Nonfinancial Firms



A: Nonfinancial Excess Bond Premium (EBP) and Nonfinancial Debt–Equity Ratio

B: Excess Bond Premium (EBP) in Financial Firms and Commercial Bank Lending Standards

Note: Figure 5 illustrates how financial distress hit the nonfinancial business sector. Figure 5A plots the debt/equity ratio of the nonfinancial corporate business sector alongside a measure of the nonfinancial excess bond premium, specifically the Gilchrist/Zakrasjek excess bond premium (EBP). Figure 5B plots the excess bond premium for financial companies (a measure of the distress the companies faced) against the tightness of bank credit (a measure derived from survey of senior loan officers about lending terms). The shaded areas mark the time from peak to trough of the Great Recession and the vertical lines mark the Lehman Brothers bankruptcy.

interest-bearing reserves. The logic for the policy was to reduce mortgage costs by expanding central bank intermediation to offset the contraction in private intermediation. Upon announcement of the program, interest rates on mortgage-backed securities fell 50 basis points and dropped another 100 as the program was phased in the following spring.

Perhaps the most dramatic intervention was the injection of equity into the commercial banking system under the Troubled Asset Relief Program (TARP), a Treasury action coordinated with the Federal Reserve in October 2008. Under the TARP, the government purchased $250 billion of preferred equity in the nine largest commercial banks. This intervention (along with temporary public guarantees on the debt of these institutions) helped replenish and stabilize the balance sheets of these institutions. In spring 2009, the Federal Reserve conducted a stress test on the commercial banks. It deemed the system as having an adequate level of capital relative to assets, marking the end of the financial crisis. The trough of the recession occurred shortly thereafter, in June 2009.

As is well-known, the recovery following the trough was quite slow. Exactly why is still a matter of debate, and we do not dig into the potential reasons in this paper. However, it is worth noting that nondurable consumption actually declines after the Lehman collapse. It is highly unusual for nondurable consumption to decline in the postwar period. As Figure 4 shows, it then remains stagnant for a long period after the trough. A number of researchers have suggested that the process of household deleveraging can help account for the slow rebound in consumption (for example, Midrigan, Jones, and Phillipon 2017).

Digging Deeper: Evidence from State Data

There has been a surge in empirical work on the issues of household balance sheets, financial frictions, and the Great Recession, often making use of cross-sectional variation. The pioneers in this area, Mian and Sufi (2014), have used regional variation to identify how the weakening of household balance sheets precipitated by the house price decline contributed to the downturn.[9] Others have focused on banks. For example, Chodorow-Reich (2014) exploits variation in bank financial health to identify how disruption in banking affected employment. Finally, there is work showing how the deterioration of nonfinancial firms' balance sheets reduced employment (for example, Giroud and Mueller 2017), again exploiting cross-sectional variation to attain identification.

In thinking about the roles of the household balance sheet channel and the disruption of financial intermediation, a natural question is whether one of these played a substantially larger role than the other in the Great Recession. Disentangling the contribution of the household balance sheet channel versus general financial market conditions on employment presents a nontrivial challenge. To date, the two phenomena have been studied separately. As we have noted, the literature on the household balance sheet channel mainly analyzes cross-sectional

[9] A few prominent examples of other papers in this vein are Kaplan, Mitman, and Violante (2017), Midrigan, Jones, and Phillippon (2017), and Berger, Guerrieri, Lorenzoni, and Vavra (forthcoming).

behavior. Conversely, work that examines the macro effects of disruptions in financial conditions (for example, Gilchrist and Zakrajšek 2012) mainly employs time series methods.

In this section, we present some evidence on this issue by examining a panel of state-level data. Following Mian and Sufi (2014) and others, we exploit the cross-sectional variation in the data to identify the effect of house prices on the regional variation in employment. We then use this information along with time series methods to disentangle the relative contributions of house prices versus disruption of intermediation to the aggregate decline in employment.

Some Patterns of Cross-Sectional and Time Series Variation

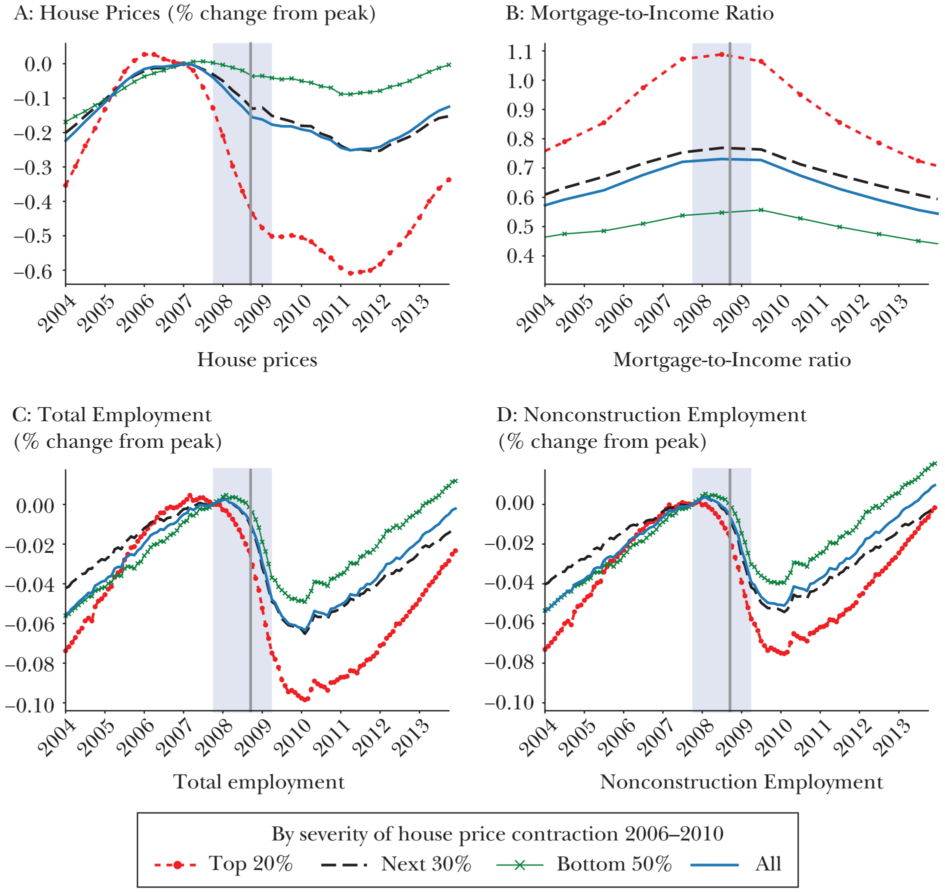
We begin with an illustration of the data before turning to our econometric framework. The panels in Figure 6 portray both the cross-sectional and time series variation of four variables: house prices, the mortgage-to-income ratio, employment, and nonconstruction employment. The data is quarterly and covers the period from 2004 to 2015. For each variable, we group states into three categories based on the severity of the house price contraction from 2006 to 2010. We then construct an aggregate of the variable for each of the three categories (the house price measure and the mortgage/income ratio are population-weighted, while the employment measures are simple aggregates). The first category of states experienced the largest house price drop. This category includes the four "sand" states—Arizona, California, Florida, and Nevada—and accounts for 20 percent of the population. Our middle group of states contains 30 percent of the population and the bottom group the remaining 50 percent. Note that our middle group has the property that it closely mirrors aggregate behavior for each variable, shown by the solid lines.

The cross-sectional patterns in the data are consistent with the evidence of the household balance sheet channel in Mian and Sufi (2014). The states experiencing the largest boom and bust in house prices also had the largest run-up in mortgage debt, as Figure 6A and B shows. In turn, there is a strong correlation between the severity of the house price decline and the corresponding employment contraction, as Figure 6C illustrates.[10] As will become clear, it is important to take into account that some of the above-average employment contraction in the sand states was the product of a collapse in residential investment as opposed to a household balance sheet channel. Construction employment fell by 40 percent in these regions. Accordingly, in Figure 6D we remove construction from the overall employment measure. The general cross-sectional relation between house prices and total employment also holds for nonconstruction employment, though with two differences. First, the

[10] As Mian and Sufi (2014) emphasize, the household balance sheet channel should affect directly nontradable employment, which depends on local demand conditions. Though we do not present the results here, we find that retail employment (their main measure of nontradable employment) exhibits the same cross-sectional correlation with house prices as total employment. In contrast, although aggregate manufacturing employment (which may be thought of as tradable goods employment) declines by 18 percent from the recession's peak to trough, there is virtually no difference in the decline across the categories of states.

Figure 6

State-Level House Prices, Mortgage Debt, and Employment



A: House Prices (% change from peak)

B: Mortgage-to-Income Ratio

C: Total Employment (% change from peak)

D: Nonconstruction Employment (% change from peak)

By severity of house price contraction 2006–2010

Note: Figure 6 portrays cross-sectional and time series variation of four variables: house prices, the mortgage-to-income ratio, employment, and nonconstruction employment. The data is quarterly and covers the period from 2004 to 2015. For each variable, we group states into three categories based on the severity of the house price contraction from 2006 to 2010. The first category experienced the largest house price drop and accounts for 20 percent of the population, the middle group contains 30 percent of the population, and the bottom group the remaining 50 percent. The solid line shows aggregate behavior for each variable. The house price and the mortgage/income ratio are population-weighted, while the employment measures are simple aggregates.

cumulative drop in nonconstruction employment is roughly 7.5 percentage points, implying that construction accounts for about 2.5 percentage points of the overall employment drop. Second, and more significant for our purposes, from early 2007 through 2008Q1, the second quarter of the recession, there is little difference in the behavior of nonconstruction employment across regions despite considerable heterogeneity in house price dynamics. The regional differences emerge later as the recession unfolds.

In addition to a clear cross-sectional pattern, our quarterly data suggests some important temporal co-movements in employment across regions. First, as we just noted, entering the business cycle peak in 2007Q4 there is a common slowdown in nonconstruction employment growth across regions that cannot be easily explained by the pattern of house price declines. As Figure 6 makes clear, the prerecession slowdown in total employment in the sand states was largely a product of the construction decline. This slowdown, however, lines up well with the unraveling of the asset-backed commercial paper market described earlier and the behavior of the various measures of financial distress plotted in Figures 4 and 5. Second, and more dramatic, around the time of the Lehman Brothers collapse, there is a rapid acceleration in the employment decline across regions. The timing of this across-the-board employment contraction mirrors the indicators of financial distress in Figure 4, which reach a peak at this point. Thus, although there are important differences across states that suggest a link between employment and house prices, there is also a considerable aggregate component to employment dynamics that is tied to economy-wide indices of financial distress.

Separating the Effects of Household Balance Sheet Stress and Financial Sector Disruption

In this section, we describe a straightforward reduced-form method to separate the effects of household balance sheets stress and financial sector disruption on the overall employment contraction during the Great Recession. To so so, we combine evidence from both cross-section and time series data. Here, we summarize the approach and the results. Details on data sources, methods, and regression results are presented in an online Appendix available with this paper at http://e-jep.org.

As our starting point, we use a panel-data vector autoregression to identify "shocks" to state-level house prices and to our indicator of aggregate financial conditions. By shocks, we mean surprise movements or "innovations" in these variables that are orthogonal to movements in employment and to each other.

For our measure of financial stress, we use the financial excess bond premium at any given time. Again, this is the spread between return on an index of financial company corporate bonds and a similar maturity government bond (after controlling for default risk). It is accordingly a measure of the external finance premium faced by financial institutions and thus a reasonable proxy for the degree of disruption of credit intermediation. As we showed in Figure 4, this premium jumps during the asset-backed commercial paper crisis and again during the Lehman fallout.

To identify shocks to the spread, we use conventional time series methods: We regress the financial excess bond premium during each time period on four lags of itself, along with current and four lags of quarterly aggregate house price growth and quarterly aggregate employment growth. The residual in this regression, call it ε_t, provides our measure of the shock to the financial excess bond premium that cannot be explained by housing prices or employment. An example of such a shock might be the jump in the spread due to the financial panic that led to the Lehman bankruptcy.

When we carry out this regression, we find that we cannot reject the hypothesis that the residuals are serially uncorrelated, implying that the estimated shocks are true surprises. This approach also makes use of timing restrictions to identify the exogenous shock ε_1 in the excess bond premium equation. In this case, given that financial markets react quickly to news, we assume that the financial excess bond premium responds immediately to current house prices and current employment growth: hence the presence of current values for those variables in the regression. However, we assume that movements in the spread affect employment and house prices only with a lag of at least one quarter, given sluggishness in the response of real sector variables to shocks. This kind of timing restriction is standard in the literature on identified vector autoregressions, but our results are robust to alternative timing assumptions.

Similarly, to obtain the shock in state-level house prices, we regress the quarterly change in house prices for each state on four lags of itself, four lags of the financial bond premium, and the current and four lagged values of that state's growth in employment. The residual in this equation $\mu_{j,t}$ provides our measure of shocks to house prices in a given state j at time t. An example of what could underlie this kind of shock is a spontaneous burst of optimism or pessimism about future house price appreciation (as in Kaplan, Mitman, and Violante 2017). This specification imposes common coefficients across states and over time, but our aggregate decomposition is insensitive to this assumption. The additional timing assumption we make in this instance is that current employment can influence housing prices, but the latter can affect the former only with a lag.

With these measures of the shocks to housing prices and financial stress in hand, our next step is to estimate the effects of these shocks on the dynamic behavior of both state-level and aggregate employment. In doing so, we interact our measures of state-level house price shocks with a state-level measure of household indebtedness. We do so in a way that permits isolating the household balance sheet channel from other ways that house prices could affect employment (for example, via the impact on residential construction). To measure the balance sheet channel, we look at the mortgage-to-income ratio in each state.

We are interested in estimating the effect of shocks to housing prices and financial intermediation over different time horizons. Having these estimates then allows us to provide a historical decomposition over the crisis period. Thus, we estimate a series of regressions with different time horizons, using state-level employment growth from one quarter up to 10 quarters ahead as the dependent variable. We include three explanatory variables in the equation below. The first variable is the shock to housing prices at the state level, $\mu_{j,t}$, taken from the earlier calculation. For the second variable, we take the mortgage-to-income ratio M_j/Y_j in a given state at the end of the house price boom, 2006Q4, which gives a sense of the vulnerability of households in that state to a decline in housing prices; we multiply by an indicator variable that takes on a value of 1 over the crisis period where house prices were declining over 2007Q1–2009Q4 and zero otherwise; and we multiply this whole term by the housing price shock at the state level. Interacting the housing price shock with the mortgage-to-income ratio provides a way to identify the balance sheet channel (analogous to Mian and Sufi 2009). Restricting the interactive effect to be operative only during the crisis captures the idea that balance sheet constraints were likely most relevant during this period.[11] The third explanatory variable is a shock to financial stress ε_1. (It is our measure of the shock to the financial excess bond premium, as described earlier in this section.)

Let $E_{j,t+h}$ be employment in state j at time $t + h$. Then the equation we estimate for the h quarter ahead growth rate of employment for each state j is given by

$$\log E_{j,t+h} - \log E_{j,t} = \beta_{p,h}\, \mu_{j,t} + \beta_{p,h}\big[Crisis = 1\big]\, \frac{M_j}{Y_j}\, \mu_{j,t} + \beta_{s,h}\ \varepsilon_t + \epsilon_{j,t,h} + \epsilon_{j,h}$$

where the regression includes a horizon-specific state fixed effect $\epsilon_{j,h}$ and an error term $\epsilon_{j,t,h}$. Note also that the coefficients are restricted to be the same across states, but are allowed to vary across the time horizon h.

Because our identified shocks to housing prices and financial conditions were obtained by conditioning on current and lagged values of state-level employment and other variables, they are orthogonal to other information that may predict future employment growth. Consequently, ordinary least squares gives consistent estimates of the coefficients. Following Jordà (2005), we can then use estimates of our equation over different horizons to construct measures of the response of employment to our identified shocks.

Table 1 reports estimates of the effect of the three explanatory variables on employment growth across horizons that span 1 to 10 quarters. The estimation period is 1992Q2 to 2015Q4. The first row of Table 1 reports the estimated effect of a house price shock over the normal course of the business cycle. The second row reports the estimated effect of a house price shock interacted with the mortgage-to-income ratio during the crisis period. The third row reports the estimated effect of a shock to financial intermediation. For all three explanatory variables, we also report the standard deviation of these estimates (in parentheses), along with the explanatory power of the regression, as measured by the R^2, at each horizon.

We find that the house price shock taken alone—that is, not operating through a balance sheet channel—has a statistically significant but modest effect at all time horizons. For example, the coefficient estimates imply that a 1 percent surprise decline in housing prices causes a 0.3 percent decline in employment growth over the next eight quarters. If we look at leverage-adjusted house price shock, which refers to the housing price shock adjusted for effects on the household balance sheet, it becomes more than twice as large as the estimated effect of house price shocks on employment during normal times. As one would expect, the employment response to these shocks varies substantially across states. For example, for states in the upper quartile of the mortgage-to-income distribution, this balance sheet

[11] As Berger, Guerrieri, Lorenzoni, and Vavra (forthcoming) argue, consumption was likely not that sensitive to house price movements during the boom phase as leverage constraints were likely not close to binding.

Table 1
Impulse Response Exercise: The Effects of Three Explanatory Variables on Employment

	Horizon									
	1	2	3	4	5	6	7	8	9	10
$\mu_{j,t}$	0.07 (0.04)	0.11 (0.05)	0.15 (0.06)	0.17 (0.07)	0.21 (0.08)	0.23 (0.09)	0.27 (0.10)	0.30 (0.12)	0.33 (0.13)	0.39 (0.15)
$\frac{M_j}{Y_j}\mu_{j,t}$	−0.10 (0.09)	−0.08 (0.14)	0.03 (0.18)	0.18 (0.21)	0.38 (0.24)	0.55 (0.25)	0.68 (0.27)	0.72 (0.29)	0.72 (0.29)	0.70 (0.31)
$\varepsilon_{j,t}$	−0.54 (0.07)	−1.14 (0.01)	−1.86 (0.12)	−2.46 (0.14)	−2.98 (0.15)	−3.48 (0.17)	−3.48 (0.18)	−3.61 (0.19)	−3.57 (0.19)	−3.62 (0.19)
R^2	0.02	0.04	0.06	0.08	0.10	0.12	0.11	0.10	0.09	0.09

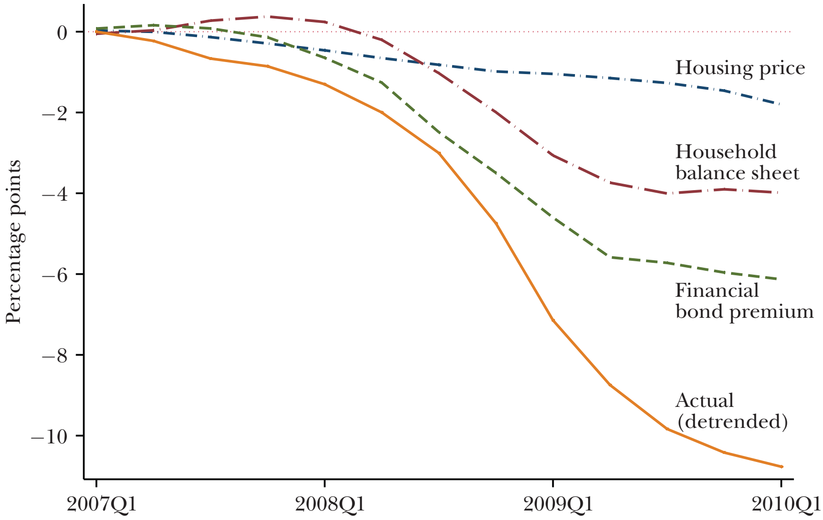
Note: Table 1 reports estimates of the effect of the three explanatory variables on employment growth across horizons that span 1 to 10 quarters. The estimation period is 1992Q2 to 2015Q4. The first row reports the estimated effect of a house price shock over the normal course of the business cycle. The second row reports the estimated effect of a house price shock interacted with the mortgage-to-income ratio during the crisis period. The third row reports the estimated effect of a shock to financial intermediation. (See text for details.) For all three explanatory variables, we also report the standard deviation of these estimates (in parentheses), along with the explanatory power of the regression, as measured by the R^2, at each horizon.

response is four times larger than the implied response for states in the lower quartile of the mortgage-to-income distribution. Interestingly, the balance sheet effect does not become economically significant until five quarters after a shock and then builds from there. This finding is consistent with the observation that differences in nonconstruction employment across states occur with a significant delay following the decline in house prices.

The estimated response of employment to a shock to the excess financial bond premium is also statistically significant and economically large. A 1 percent surprise increase in the excess financial bond premium implies a 3.6 percent drop in employment at the two-year horizon. These estimates are comparable to those obtained by Gilchrist and Zakrajšek (2012) using a standard vector autoregression methodology to compute impulse responses.

Armed with the estimates of the effects of the various shocks, our final step is to construct a historical decomposition that measures the relative contribution of housing price changes, the household balance sheet channel, and the deterioration in overall financial conditions to the decline in aggregate employment that occurred over the 2007–2010 period. We first take the estimated effects in Table 1 and multiply them by the relevant shocks obtained from our forecasting equations. We then compute the weighted sum of these effects across states to obtain the impact of a shock that occurs in a given time period on *h*-period ahead growth in aggregate employment growth. Because the shocks are serially uncorrelated, we can sum up the estimated effect of each historical shock at each horizon to obtain the total response of employment to the past history of shocks.

Figure 7
Employment Decomposition by Type of Shock



Note: Figure 7 displays the cumulative contribution of each of three shocks (housing price, household balance sheet, and financial bond premium shocks) to aggregate employment over the period 2007Q1 to 2010Q1 along with the realized path of aggregate employment (measured as a deviation from a linear trend).

Figure 7 displays the cumulative contribution of each of these shocks (housing price, household balance sheet, and financial bond premium shocks) to aggregate employment over the period 2007Q1 to 2010Q1 along with the realized path of aggregate employment (measured as a deviation from a linear trend). Aggregate employment fell by 9 percentage points relative to trend over this time period. The linear effect of house price shocks on aggregate employment is modest and implies a 1.7 percent decline in employment over this time period. In contrast, the household balance sheet effect estimated during the crisis is sizeable and implies a 4.1 percent decline in aggregate employment. The shock to the financial bond premium provides the largest effect however and explains a 5.7 percentage point decline in employment during this period. Notably, the shock to the financial bond premium that occurred during the 2008Q3 Lehman collapse accounts for 3.5 percentage points of the overall employment contraction. In contrast, the Lehman collapse explains none of the decline in employment associated with house prices or household balance sheets. Thus, although the direct effect of house prices on household balance sheets is an important component of the decline in aggregate output, our estimates imply that the recession would have been far milder in the absence of the financial turmoil that ensued.[12]

[12] We note that our estimate of the effect of the financial shock on employment is conservative in the sense that we do not allow shock to the excess bond premium to affect current house prices but do let the

We conclude with two qualifications for this exercise. First, it is important to emphasize the reduced-form nature of our exercise. It is reasonable to argue that the house price shock interacted with leverage captures the household balance sheet channel and that the shock to the financial excess bond premium captures the disruption of intermediation. However, the cumulative impact on employment depends on all the propagation mechanisms that are operative. For example, the weakening of the economy in response to either shock can give rise to tightening of financial constraints on nonfinancial firms, as we argued earlier. What this suggests is that a full accounting of how the financial crisis played out will require structural modeling.

Second, we identify orthogonal shocks to house prices and credit spreads by using a linear vector autoregression in conjunction with restrictions on their contemporaneous interaction. However, the large jumps in the financial excess bond premium plotted in Figures 4 and 5 likely have their origins in house price declines that led to mortgage defaults, which in turn unsettled financial markets. This phenomenon may not be well-captured in a linear regression. However, by including housing prices interacted with leverage, we have controlled for the main way that housing prices could have a nonlinear effect on employment independent of financial market disruption. It is thus reasonable to treat our identified shock to the credit spread as exogenous from the standpoint of identifying its effects on employment. Nonetheless, incorporating nonlinearities explicitly in the estimation would be desirable. Again, this would likely involve a more structural approach.

Conclusion

Gaining a deeper understanding of the Great Recession is important, because the lessons that arise from that event will shape our perceptions of how the macroeconomy works, and sometimes doesn't work, for years to come. We have argued on theoretical and empirical grounds that financial distress in each of the three main sectors—households, financial intermediaries, and nonfinancial firms—played a meaningful role in the evolution of the Great Recession. Our empirical exercise suggests that while the household balance sheet channel and the disruption of financial intermediation contributed significantly to the overall employment contraction, the recent recession would have been relatively mild without the disruption of financial intermediation.

Of course, understanding the Great Recession ultimately requires more than looking at the downturn. We also need a better understanding of the run-up to the crisis and the slow recovery afterward. For example, purely fundamentals-based models have difficulty accounting for the boom and then subsequent bust in house prices. This opens up the possibility for a behavioral approach to explain how a wave

latter affect the former. Under the alternative extreme, where the bond premium shock affects current house prices but not the reverse, the financial shock explains a 6.4 percent employment decline while the leveraged adjusted house price shock accounts for 3.7 percent.

of optimism turned to pessimism in housing markets, though a widely accepted approach along these lines has yet to materialize. For the slow recovery, we know from Reinhart and Rogoff (2009) that recoveries from financial crises are often much longer than normal. Although broad measures of financial stress suggest that financial markets normalized to a considerable extent by 2009, there is some evidence that tightness in credit markets persisted for both households (Midrigan, Jones, and Phillipon 2017) and small businesses (Chen, Hanson, and Stein 2017). Accounting for the slow recovery, including the role of financial factors, is an important topic for future research.

■ *Thanks to Ben Bernanke for helpful comments and to Sneha Agrawal and James Graham for excellent research assistance.*

References

Adrian, Tobias, Paolo Colla, and Hyun Song Shin. 2013. "Which Financial Frictions: Parsing the Evidence from the Financial Crisis of 2007 to 2009." In *NBER Macroeconomics Annual 2012*, vol. 27, edited by Daron Acemoglu, Jonathan Parker, and Michael Woodford, 159–214. University of Chicago Press.

Adrian, Tobias, and Hyun Song Shin. 2010. "The Changing Nature of Financial Intermediation and Financial Crisis of 2007–2009." *Annual Review of Economics* 2: 603–618.

Arellano, Cristina, Yan Bai, and Patrick Kehoe. 2016. "Financial Frictions and Fluctuations in Volatility." NBER Working Paper 22990.

Becker, Bo, and Victoria Ivashina. 2014. "Cyclicality of Credit Supply: Firm Level Evidence." *Journal of Monetary Economics* 62: 76–93.

Benmelech, Efraim, Ralf R. Meisenzahl, and Rodney Ramacharan. 2017. "The Real Effects of Liquidity During the Financial Crisis: Evidence from Automobiles." *Quarterly Journal of Economics* 132(1): 317–65.

Berger, David, Veronica Guerrieri, Guido Lorenzoni, and Joseph Vavra. Forthcoming. "House Prices and Consumer Spending." *Review of Economic Studies*.

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

Bernanke, Ben S. 2010. "Causes of the Recent Financial and Economic Crisis." Statement before the Financial Crisis Inquiry Commission, Washington, DC, September 2. https://www.federalreserve.gov/newsevents/testimony/bernanke20100902a.htm.

Bernanke, Ben S. 2015. *The Courage to Act: A Memoir of a Crisis and Its Aftermath.* New York: Norton.

Bernanke, Ben, and Mark Gertler. 1989. "Agency Costs, Net Worth, and Business Fluctuations." *American Economic Review* 79(1): 14–31.

Bernanke, Ben, and Mark Gertler. 1995. "Inside the Black Box: The Credit Channel of Monetary Policy Transmission." *Journal of Economic Perspectives* 9(4): 27–48.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." Chap. 21 in *Handbook of Macroeconomics*, vol. 1, edited by Ben Friedman and Michael Woodford. Elsevier.

Brunnermeier, Markus K., Thomas M. Eisenbach, and Yuliy Sannikov. 2013. "Macroeconomics with Financial Frictions: A Survey." In *Advances in Economics and Econometrics: Tenth World Congress of the Econometric Society*, vol. 2: *Applied Economics*, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 3–94. Cambridge University Press.

Brunnermeier, Markus K., and Yuliy Sannikov. 2014. "A Macroeconomic Model with a Financial Sector." *American Economic Review* 104(2): 379–421.

Chen, Brian S., Samuel G. Hanson, and Jeremy C. Stein. 2017. "The Decline of Big-Bank Lending to Small Businesses: Impacts on Local Credit and Labor Markets." NBER Working Paper 23843.

Chodorow-Reich, Gabriel. 2014. "The Employment Effects of Credit Market Disruptions: Firm-level Evidence from the 2008–9 Financial Crisis." *Quarterly Journal of Economics* 129(1): 1–59.

Chodorow-Reich, Gabriel, and Antonio Falato. 2017. "The Loan Covenant Channel: How Bank Health Transmits to the Real Economy." NBER Working Paper 23879.

Christiano, Lawrence J., Roberto Motto, and Massimo Rostagno. 2014. "Risk Shocks." *American Economic Review* 104(1): 27–65.

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

Cúrdia, Vasco, and Michael Woodford. 2011. "The Central-Bank Balance Sheet as an Instrument of Monetary Policy." *Journal of Monetary Economics* 58(1): 54–79.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91(3): 401–419.

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

Eggertsson, Gauti B., and Paul Krugman. 2012. "Debt, Deleveraging, and Liquidity Trap: A Fisher–Minsky–Koo Approach." *Quarterly Journal of Economics* 127(3): 1469–1513.

Fisher, Irving. 1933. "The Debt-Deflation Theory of Great Depressions." *Econometrica* 1(4): 337–57.

Geanakopolos, John. 2010. "The Leverage Cycle." Chap. 1 in *NBER Macroeconomics Annual 2009*, vol, 24, edited by Daron Acemoglu, Kenneth Rogoff, and Michael Woodford. University of Chicago Press.

Gertler, Mark, and Peter Karadi. 2011. "A Model of Unconventional Monetary Policy." *Journal of Monetary Economics* 58(1): 17–34.

Gertler, Mark and Nobuhiro Kiyotaki. 2015. "Banking, Liquidity and Bank Runs in an Infinite Horizon Economy." *American Economic Review* 105(7) 2011–43.

Gertler, Mark, and Nobuhiro Kiyotaki. 2011. "Financial Intermediation and Credit Policy in Business Cycle Analysis." Chap. 11 in *Handbook of Monetary Economics*, vol. 3A, edited by Benjamin M. Friedman, and Michael Woodford. Amsterdam, Netherlands: Elsevier.

Gertler, Mark, Nobuhiro Kiyotaki, and Andrea Prestipino. 2016. "Wholesale Banking and Bank Runs in Macroeconomic Modeling of Financial Crises." Chap 16 in *Handbook of Macroeconomics*, vol. 2, edited by John B. Taylor and Harald Uhlig.

Gertler, Mark, Nobuhiro Kiyotaki, and Andrea Prestipino. 2017. "A Macroeconomic Model with Financial Panics." NBER Working Paper 24126.

Gilchrist, Simon, Raphael Schoenle, Jae Sim, and Egon Zakrajšek. 2017. "Inflation Dynamics During the Financial Crisis." *American Economic Review* 107(3): 785–823.

Gilchrist Simon, Michael Siemer and Egon Zakrajšek. 2017. "The Real Effects of Credit Booms and Busts: A County-Level Analysis." Unpublished paper.

Gilchrist, Simon, and Egon Zakrajšek. 2012. "Credit Spreads and Business Cycle Fluctuations." *American Economic Review* 102(4): 1692–1720.

Giroud, Xavier, and Holger Mueller. 2017. "Firm Leverage, Consumer Demand, and Unemployment During the Great Recession." *Quarterly Journal of Economics* 132(1): 271–316.

Gorton, Gary B. 2010. *Slapped by the Invisible Hand: The Panic of 2007.* Oxford University Press.

Guerrieri, Veronica, and Guido Lorenzoni. 2017. "Credit Crises, Precautionary Saving and the Liquidity Trap." *Quarterly Journal of Economics* 132(3): 1427–67.

He, Zhiguo, and Arvind Krishnamurthy. 2013. "Intermediary Asset Pricing." *American Economic Review* 103(2): 732–70.

He, Zhiguo, and Arvind Krishnamurthy. 2014. "A Macroeconomic Framework for Quantifying Systemic Risk." NBER Working Paper 19885.

Iacoviello, Matteo. 2005. "House Prices, Borrowing Constraints and Monetary Policy." *American Economic Review* 95(3): 739–64.

Jermann, Urban, and Vincenzo Quadrini. 2012. "Macroeconomic Effects of Financial Shocks." *American Economic Review* 102(1): 238–71.

Jordà, Òscar. 2005. "Estimation and Inference of Impulse Responses by Local Projections." *American Economic Review* 95(1): 161–82.

Justiniano, Alejandro, Giorgio E. Primiceri, and Andrea Tambalotti. 2010. "Investment Shocks and Business Cycles." *Journal of Monetary Economics* 57(2): 132–45.

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

Kaminsky, Graciela L., and Carmen M. Reinhart. 1999. "The Twin Crises: The Causes of Banking and Balance of Payments Problems."

American Economic Review 89(3): 473–500.

Kaplan, Greg, Kurt Mitman, and Gianluca L. Violante. 2017. "The Housing Boom and Bust: Model Meets Evidence." NBER Working Paper 23694.

Kashyap, Anil K., Jeremy C. Stein, and David W. Wilcox. 1993. "Monetary Policy and Credit Conditions: Evidence from the Composition of External Finance." *American Economic Review* 83(1): 78–98.

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

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

Mendoza, Enrique G. 2010. "Sudden Stops, Financial Crises, and Leverage." *American Economic Review* 100(5): 1941–66.

Mian, Atif, Kamalesh Rao, and Amir Sufi. 2013. "Household Balance Sheets, Consumption, and the Economic Slump." *Quarterly Journal of Economics* 128(4): 1687–1726.

Mian, Atif and Amir Sufi. 2014. "What Explains the 2007–2009 Drop in Employment." *Econometrica* 82(6): 2197–2223.

Mian, Atif, Amir Sufi, and Emil Verner. 2017. "Household Debt and Business Cycles Worldwide." *Quarterly Journal of Economics* 132(4): 1755–1817.

Midrigan, Virgiliu, Callum Jones, and Thomas Phillipon. 2017. "Household Leverage and the Recession." Unpublished paper.

Reinhart, Carmen M., and Kenneth S. Rogo. 2009. *This Time is Different: Eight Centuries of Financial Folly.* Princeton.

Schularick, Moritz, and Alan M. Taylor. 2012. "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870–2008." *American Economic Review* 102(2): 1029–61.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 31–58

Finance and Business Cycles: The Credit-Driven Household Demand Channel

Atif Mian and Amir Sufi

What is the role of the financial sector in explaining business cycles? This question is as old as the field of macroeconomics, and an extensive body of research conducted since the global financial crisis of 2008 has offered new answers. The specific idea put forward in this article is that expansions in credit supply, operating primarily through household demand, have been an important driver of business cycles. We call this the credit-driven household demand channel. While this channel helps explain the recent global recession, it also describes economic cycles in many countries over the past 40 years.

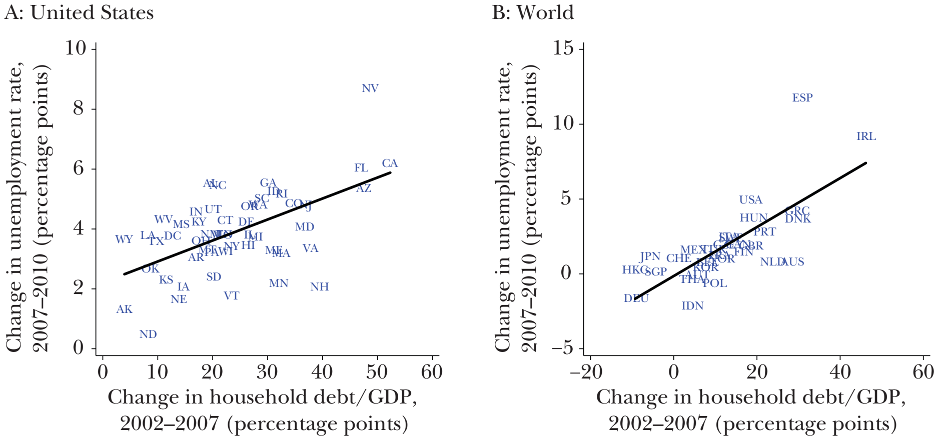
Our interest in this topic began with a striking empirical regularity of the Great Recession: the larger the increase in household leverage prior to the recession, the more severe the subsequent recession. Figure 1 shows this pattern both across states within the United States and across countries in the world. Indeed, the ability of household debt expansion to predict recession severity across geographical areas during the Great Recession has been demonstrated by a number of studies (for example, Mian and Sufi 2010; Glick and Lansing 2010; IMF 2012; Martin and Philippon 2017). The ability of household debt expansion to predict a slowdown in growth is broader than the Great Recession. A rise in household debt is a robust

■ *Atif Mian is John H. Laporte, Jr. Class of 1967 Professor of Economics and Public Affairs, Woodrow Wilson School of Public Affairs, Princeton University, Princeton, New Jersey. Amir Sufi is Bruce Lindsay Professor of Economics and Public Policy, Booth School of Business, University of Chicago, Chicago, Illinois. Both authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts.*

[†] For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at

https://doi.org/10.1257/jep.32.3.31 doi=10.1257/jep.32.3.31

Figure 1
Household Debt and Unemployment



Note: Figure 1 shows the relationship between change in household leverage and change in unemployment rate, both across states within the United States and across countries in the world.

predictor of a decline in GDP growth across a large number of countries since the 1970s.

This empirical pattern can be explained by the credit-driven household demand channel, which rests on three pillars. First, an expansion in credit supply, as opposed to technology shocks or permanent income shocks, is a key force generating expansion and contraction in economic activity. Second, the expansionary phase of the credit cycle affects the real economy primarily by boosting household demand as opposed to boosting productive capacity of firms in the economy. Third, the contraction in the aftermath of a large increase in household debt is driven initially by a decline in aggregate demand, which is amplified by nominal rigidities, constraints on monetary policy, and banking sector disruptions.

The contractionary phase of the business cycle is a consequence of the excesses generated during the expansionary phase; financial crises and a sudden collapse in credit supply are not exogenous events hitting a stable economy. As a result, we must understand the boom to make sense of the bust. Our emphasis on the relationship between expansion and contraction is reminiscent of the perspective taken by earlier scholars such as Kindleberger (1978) and Minsky (2008). We discuss how the presence of behavioral biases and aggregate demand externalities may be able to generate endogenous boom-bust credit cycles.

What triggers the expansion in credit supply that initiates the credit cycle and its business cycle implications? Answers to this question are less definitive at this point. Based on an analysis of historical episodes, we conclude that a shock that leads to a rapid influx of capital into the financial system often triggers an expansion in credit supply. Recent manifestations of such a shock are the rise in income inequality in the United States (Kumhof, Rancière, and Winant 2015) and the rapid rise in savings by many emerging markets (that is, the "global savings glut" as articulated by Bernanke 2005).

The discussion of fundamental causes of credit supply expansion naturally leads to consideration of longer-run factors. For example, there has been a long-term secular rise in private credit-to-GDP ratios, especially household credit-to-GDP ratios (Jordà, Schularick, and Taylor 2016). This rise has been accompanied by a decline in real long-term interest rates, and a rise in within-country inequality and across-country "savings gluts." There may be a connection between these longer-term trends and what we uncover at the business cycle frequency. We discuss these issues in the conclusion.

Credit Supply Expansion and Business Cycles

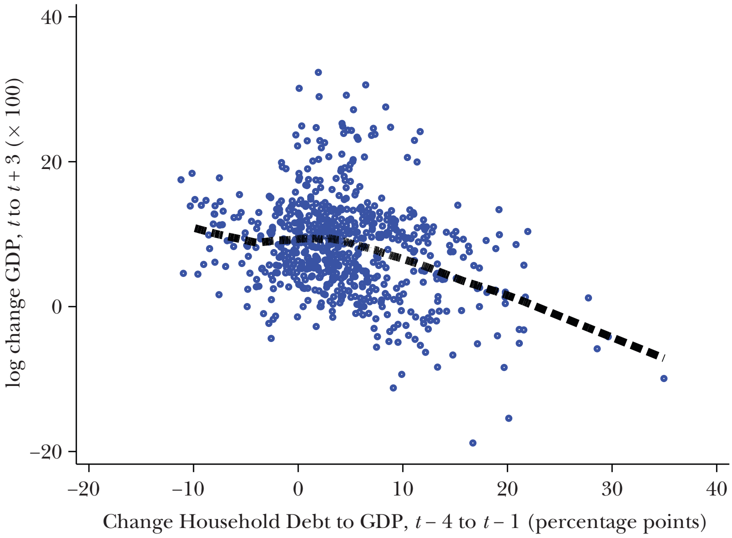
Credit Cycles and Business Cycles

A robust empirical finding is the existence of predictable credit cycles that generate fluctuations in real economic activity. López-Salido, Stein, and Zakrajšek (2017) present evidence on the predictability of the credit cycle; they use evidence from the United States since the 1920s to show that a narrowing of the spread between mid-grade corporate bonds and US Treasuries predicts a subsequent widening of credit spreads. Krishnamurthy and Muir (2017) use a sample of 19 countries (with data going back to the 19th century for 14 countries) to show that a period of low credit spreads precedes a sudden widening of credit spreads. The notion of a predictable cycle in credit is also highlighted by Borio (2014), who reviews a substantial body of research from the Bank of International Settlements supporting this view.

This predictable cycle has important effects on the household debt cycle. Using a sample of 30 mostly advanced countries over the last 40 years, in Mian, Sufi, and Verner (2017b), we estimate a vector autogression in the level of household debt to lagged GDP, nonfinancial firm debt to lagged GDP, and log real GDP. The results show that a sudden increase in the household debt-to-GDP ratio in a given country leads to a three-year increase in the household debt-to-GDP ratio followed by a sharp fall over the subsequent seven years. There is a predictable decline in household debt following a positive shock to household debt in a country, which reflects the importance of the predictable cycle in credit spreads. The household debt cycle is closely connected to the business cycle. We show that a shock to household debt generates a boom-bust cycle in the real economy that is similar to the credit cycle. Growth increases for two to three years, and then falls significantly.

The International Monetary Fund (2017) estimates similar specifications using a significantly larger sample of 80 countries, with some data going back to the 1950s. This work confirms the boom-bust pattern associated with sudden increases in the household debt-to-GDP ratio. The report concludes that "an increase in household debt boosts growth in the short-term but may give rise to macroeconomic and financial stability risks in the medium term." Their sample includes substantially

Figure 2
Rise in Household Debt Predicts Lower GDP Growth



Note: Figure 2 is based on a sample of 30 mostly high-income countries from 1960 to 2012 in the Mian, Sufi, and Verner (2017b) sample. Each point represents a given country and a given year. This figure plots real GDP growth from year t to $t + 3$ against the rise in the household debt to GDP ratio from year $t - 4$ to year $t - 1$. See Mian, Sufi, and Verner (2017b) for more details.

more emerging economies, and they are able to show that the same pattern is present in emerging economies, but it is less pronounced. Drehmann, Juselius, and Korinek (2017) also confirm this pattern in a panel of 17 advanced economies from 1980 to 2015. They emphasize the importance of rising debt service burdens in explaining the subsequent drop in GDP.

In short, a boom-bust cycle of credit and housing debt is a robust pattern in the data. The pattern is strong enough that a rise in household debt systematically predicts a decline in subsequent GDP growth. Figure 2 is based on a sample of 30 mostly high-income countries from 1960 to 2012 in the Mian, Sufi, and Verner (2017b) sample. Each point represents a given country and a given year: for example, the point to the farthest right shown is for Ireland in 2007. This data point shows the change in household debt in Ireland from 2003 to 2006 (shown on the horizontal axis) is associated with a large decline GDP for Ireland from 2007 to 2010 (shown on the vertical axis). The dotted line is a nonparametric plot of the relationship in the data. Overall, there is a robust negative correlation between the growth in household debt from $t - 4$ to $t - 1$ and the subsequent real GDP growth from t to $t + 3$.

Intriguingly, professional forecasters do not seem to take into account the connection between increases in household debt and lower subsequent growth. In

Mian, Sufi, and Verner (2017b), we examine output growth forecasts by the IMF and the OECD and find that growth is systematically over-forecasted following periods of high household debt.

Identification of Credit Supply Expansion in Aggregate Data

Why might household debt increase suddenly? Why might such a rise generate a boom-bust cycle in real economic activity? An initial approach to answer this question focuses on whether debt expansion is due to credit demand shocks or credit supply shocks. By credit demand shocks, we mean changes in household permanent income, demographics, or beliefs. By credit supply shocks, we mean an increased willingness of lenders to provide credit that is independent of the borrower's income position.[1]

Two approaches have been used to distinguish credit supply shocks from credit demand shocks. In this subsection, we look at aggregate country-level analysis in datasets that cover a long time series and many macroeconomic cycles. In the next subsection, we look at studies that focus on specific macroeconomic episodes and use cross-sectional data across countries or regions.

When using longer time series datasets covering many episodes, the most direct empirical method for separating credit supply versus credit demand shocks is to examine interest rates and credit spreads during household debt expansions. Such evidence favors the credit supply expansion view. For example, using the same sample as in Figure 2, in Mian, Sufi, and Verner (2017b), we show that large three- to four-year increases in household debt are associated with low spreads between mortgage credit and sovereign credit. To isolate increases in credit supply, we use episodes in which mortgage credit spreads are low as an instrument for the rise in household debt, and we show that such credit-supply-driven increases in household debt predict subsequent economic downturns.[2]

In another approach, Jordà, Schularick, and Taylor (2015) use pegged currencies and monetary policy shocks to isolate variation in credit supply. Countries with fixed exchange rates see changes in short-term interest rates that are unrelated to home economic conditions when monetary policy shifts in the pegged country. They show that monetary policy shocks that lower the short-term interest rate are associated with an increase in household debt and house prices. Furthermore, the rise in household debt and house prices heightens the risk of a financial crisis.

[1] We note from the outset that the negative relationship between a rise in household debt and subsequent growth shown in Mian, Sufi, and Verner (2017b) casts doubt on the role of credit demand shocks coming from changes in permanent income. A rise in household debt driven by a positive permanent income shock should predict an increase in subsequent growth, at least in models with rational expectations. The opposite pattern is found in the data.

[2] More specifically, we present the impulse response function of log real GDP to an increase in household debt from a proxy structural vector autoregression specification in which low credit spread episodes are used as an instrument for a credit-supply-driven increase in household debt. The proxy structural vector auto regression approach is based on Mertens and Ravn (2013); see Mian, Sufi, and Verner (2017b) for details. The impulse response function from this specification shows a similar boom and bust in real economic activity coming from a credit-supply-driven increase in household debt.

Krishnamurthy and Muir (2017) find similar results when examining growth in private sector credit. In their sample of 19 countries going back to the 19th century, they show that credit spreads between lower- and higher-grade bonds within a country tend to fall in the period of credit growth that occurs before a financial crisis. They conclude based on the evidence that the "behavior of both prices and quantities suggests that credit supply expansions are a precursor to crises."

Identification of Credit Supply Expansion in Specific Episodes

A study of specific episodes can allow for a cleaner identification of credit supply shocks. Here, we focus on three types of episodes: US banking deregulation episodes, the introduction of the euro in the early 2000s, and US credit standards in the lead-up to the Great Recession.

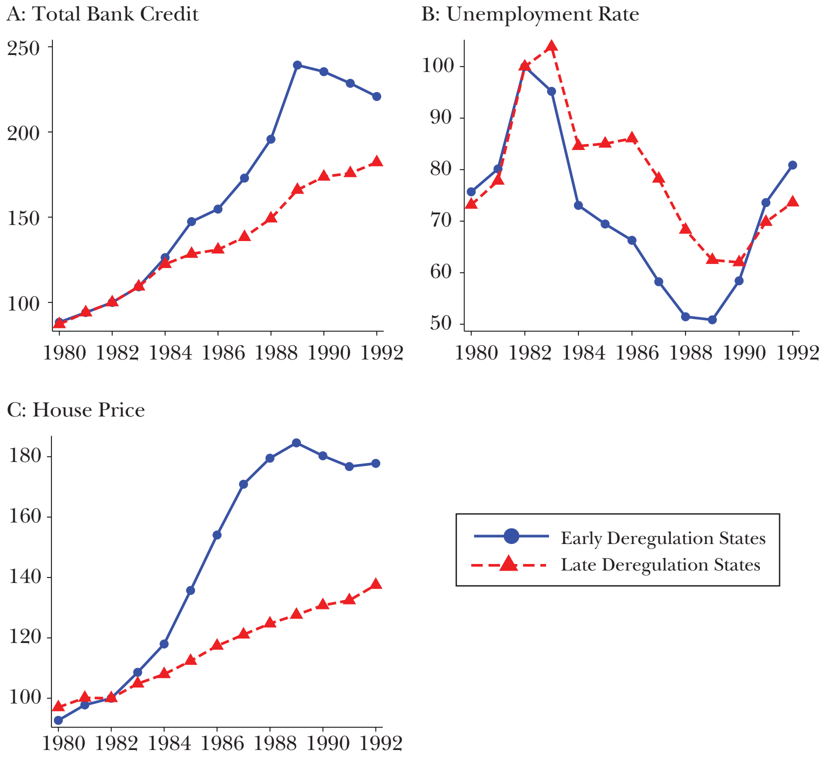
Perhaps the cleanest identification of credit supply shocks in recent literature comes from the evaluation of banking deregulation episodes. Di Maggio and Kermani (2017) focus on the federal preemption of state laws against predatory lending. As of January 1, 2004, 18 US states had anti-predatory lending rules that applied to all banks doing business in the state. However, the Office of the Comptroller of the Currency adopted regulations preempting these state laws from applying to national banks. They show that the states where anti-predatory-lending rules were preempted witnessed a surge in mortgage credit provided by national banks in 2005 and 2006, which corresponded to a sudden increase in house prices and employment in the nontradable sector. The same states then witnessed a larger decline in house prices, mortgage availability, and employment in the nontradable sector from 2006 to 2010. The preemption of anti-predatory lending laws apparently induced a credit-supply-driven boom and bust.

In a study of an earlier episode, in Mian, Sufi, and Verner (2017a), we focus on US banking deregulation in the 1980s. We classify a state as more deregulated as of 1983 if the state was early to remove restrictions on inter- and intrastate bank branching (for example, Jayaratne and Strahan 1996; Kroszner and Strahan 2014). As Figure 3A shows, states that deregulated their banking system earlier witnessed a larger rise in bank credit from 1983 to 1989 relative to states that deregulated their banking system later. During the expansion period, the unemployment rate fell by more in early deregulation states (Figure 3B) and house price growth was significantly stronger in these states (Figure 3C). After the recession hit in 1990, relative to late deregulation states, early deregulation states saw a rise in the unemployment rate and a decline in house prices. States with a stronger credit supply shock from 1983 to 1989 experienced a significantly amplified business cycle.

These two studies both build on a precise source of variation in bank regulation to generate differential credit supply expansion across states, and they find similar results: stronger credit supply expansion due to a different bank regulatory environment generates stronger growth in debt in the short run and a more severe recession in the medium run.

The introduction of the euro currency in the late 1990s can be viewed as a shock that increased credit supply by lowering currency and other risk premia,

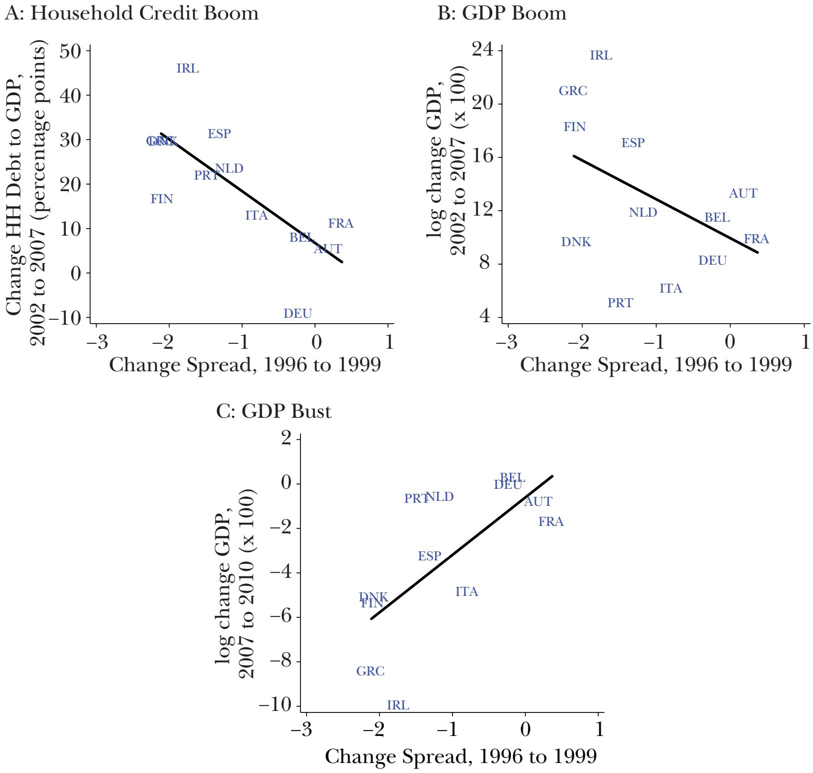
Figure 3
US Banking Deregulation Quasi-Experiment
(outcomes indexed to 100 in 1982)



Note: We focus on US banking deregulation in the 1980s and classify a state as more deregulated as of 1983 if the state was early to remove restrictions on inter- and intrastate bank branching. This figure plots outcomes for states that deregulated early versus late. For more information, see Mian, Sufi, and Verner (2017a).

especially in peripheral European countries (Mian, Sufi, and Verner 2017b). The decline in risk premia for a given country can be most easily seen in the spread between interest rates on sovereign bonds in the country versus US bonds. For example, Denmark, Finland, Ireland, and Greece all witnessed substantial declines in their borrowing costs on sovereign debt relative to US Treasury rates. Figure 4A shows that countries with the largest decline in this interest spread from 1996 to 1999 experienced the largest increase in household debt from 2002 to 2007. Figure 4B shows that countries seeing the biggest drop in this interest spread also see the strongest GDP growth from 2002 to 2007. Figure 4C shows that these same countries experienced a worse economic downturn from 2007 to 2010. We interpret this evidence as showing how a credit supply expansion induced by an institutional change led to a boom in household debt and in the real economy, followed by a more severe economic downturn.

Figure 4
Eurozone Quasi-Experiment


A: Household Credit Boom
B: GDP Boom
C: GDP Bust

Note: This figure plots various outcomes against the change in the sovereign interest spread from 1996 to 1999 in countries that joined the euro currency zone. The sovereign interest spread is the interest rate on the 10-year government bond of the given country relative to the interest rate on the 10-year government bond of the United States. Please see Mian, Sufi, and Verner (2017b) for more details.

An alternative measure of shifts in European credit supply are credit standards as reported by loan officers at banks (for example, Favilukis, Kohn, Ludvigson, and Van Nieuwerburgh 2012). The European Central Bank carried out a survey of loan officers across the euro area starting in 2003, asking: "Over the past three months, how have your bank's credit standards as applied to the approval of loans to households changed?" The survey indicates that the credit expansion period of 2003 to 2007 was associated with a substantial loosening of credit standards by loan officers on house purchase loans, especially in late 2004 and 2005.

The rapid increase in household debt in the United States from 2000 to 2007 has been studied extensively, and many factors, including credit spreads, loan surveys, and the innovation of private-label securitization, all point to an expansion in credit supply. Risk spreads on mortgage credit fell sharply from 2000 to 2005 (for example, Pennington-Cross and Chomsisengphet 2007; Demyanyk and Van Hemert 2011). Justiniano, Primiceri, and Tambalotti (2017) point to a "mortgage rate conundrum" in the summer of 2003 when mortgage credit spreads relative to US Treasuries fell 80 basis points, and then continued to fall through 2005. Evidence on credit standards in the United States points in the same direction. According to the Federal Reserve Board Senior Loan Officer Opinion Survey, the loosening of credit standards on US mortgages is remarkably similar to the European pattern examined in Favilukis et al. (2012).

The shift in US credit supply can also be seen in the dramatic changes in mortgage markets during the late 1990s and early 2000s. Levitin and Wachter (2012) conduct a detailed analysis of the rise of the private-label securitization market, which increased from about 15 percent of all mortgage originations to almost 50 percent in 2004 and 2005. Private-label securitization refers to mortgages that were neither retained by the bank issuing the mortgage, nor issued by a government-sponsored enterprise like Fannie Mae or Freddie Mac. The rise of the private-label securitization market was accompanied by a rise in subprime mortgages, which together represented a positive credit supply shock to marginal borrowers who were previously denied credit (for example, Mayer 2011; Mian and Sufi 2009; Demyanyk and Van Hemert 2011). In particular, a rise in securitized subprime mortgages reduced the incentives of financial intermediaries to screen borrowers, thereby helping to explain why default rates on these mortgages were so high (Keys, Mukherjee, Seru, and Vig 2010). Fraud was rampant in the private-label securitization market during the height of the mortgage credit boom (Piskorski, Seru, and Witkin 2015; Griffin and Maturana 2016b; Mian and Sufi 2017a), which likely helped fuel house price growth in some areas of the country (Griffin and Maturana 2016a).

This is not to say that the subprime mortgage market alone can explain the sharp rise in household debt in the United States from 2000 to 2007. Borrowing by existing homeowners was an important driver of aggregate household debt, and such borrowing occurred even among higher credit score borrowers (for example, Mian and Sufi 2011; Mian, Sufi, and Verner 2017a). Indeed, there was an expansion in credit supply from 2001 to 2005 across the credit score distribution (Anenberg, Hizmo, Kung, and Molloy 2017).

Credit Supply Expansion and House Prices

The interaction between house prices and credit supply expansions has led to the question of whether the increase in house prices is the initial shock and the rise in household debt is a response, as argued by Laibson and Mollerstrom (2010), Foote, Girardi, and Willen (2012), and Adelino, Schoar, and Severino (2017). For example, it could have been that an "optimism" shock led to a rise in house prices, and credit merely followed the rise in house prices. There are no doubt feedback effects between the housing market and credit supply expansions. For example, an initial expansion in credit supply may lead to a rise in house prices, thereby boosting residential investment and encouraging lenders to provide even more credit because they expect house prices to rise further.

However, the weight of the empirical evidence suggests that house prices are more likely to be a *response* to credit supply expansion rather than a *cause*. A substantial body of research using careful identification strategies in microeconomic settings shows that exogenous changes in credit supply have quantitatively large effects on house prices (for example, Adelino, Schoar, and Severino 2014; Favara and Imbs 2015; Di Maggio and Kermani 2017; Mian, Sufi, and Verner 2017a). There is also a body of research using quantitative macroeconomic models to show how changes in credit affect house prices (for example, Favilukis, Ludvigson, and Nieuwerburgh 2017; Justiniano, Primiceri, and Tambalotti 2015; Landvoigt 2016).

Country-level datasets also support the view that credit supply initiates the rise in house prices. In the study mentioned earlier of survey data from loan officers, Favilukis et al. (2012) use credit standards data for the 2002 to 2010 period for 11 countries, including the United States and a panel of European economies. They conclude that "a stark shift in bank lending practices ... was at the root of the housing crisis." Using the sample of 30 countries over the past 40 years, in Mian, Sufi, and Verner (2017b), we run a bivariate recursive vector autoregression to examine the dynamic relationship between increases in household debt and house prices. We find that a shock to household debt leads to a large and immediate increase in house prices, followed by substantial mean reversion four years after the initial shock. In contrast, a shock to house price growth leads to a gradual rise in household debt to a permanently higher level, but not to any boom and bust dynamics. For further discussion of the relative importance of credit supply expansion versus a rise in house prices, see Mian and Sufi (2017b).

The rise in house prices driven by credit supply expansion is of central importance for the aggregate economy, as it boosts construction activity, retail employment, and consumption. In addition, the rise in house prices is an amplification mechanism because households often borrow aggressively against the rising value of their home (Mian and Sufi 2011). Many of these real effects help explain the severity of the subsequent downturn, and we return to these issues later in this paper.

The Household Demand Channel

Credit supply expansions generate a boom-bust cycle in real economic activity. But what is the precise channel? An expansion in credit supply could affect the supply side of the economy by boosting firm investment or employment. Alternatively, it could boost aggregate demand by enabling households to increase consumption. There are good theoretical arguments for why credit supply could operate through the firm or household channel, and there are certainly episodes in history where credit supply boosted the economy through the firm sector. However, in recent history, the household demand channel appears dominant.

For example, over the past 40 years, the boom-bust business cycle generated by a rise in debt is unique to household debt; increases in firm debt or government debt do not produce the same pattern (Mian, Sufi, and Verner 2017b). Furthermore, periods of rising household debt are associated with a rise in the consumption-to-GDP ratio, an increase in imports of consumption goods, and no change in the investment-to-GDP ratio. In advanced economies, a rise in household debt generates a consumption boom-bust cycle that is significantly more severe than the real GDP boom-bust cycle (IMF 2017). Household debt appears to be crucial in generating these cycles; for example, a rise in the consumption-to-GDP ratio by itself does not predict subsequently lower growth. But a rise in consumption-to-GDP ratios concurrent with a large rise in household debt does predict lower growth (Mian, Sufi, and Verner 2017b).

Household debt also appears to be important in predicting financial crises. Jordà, Schularick, and Taylor (2016) use their disaggregated bank credit dataset to estimate the relationship between bank credit and subsequent financial crises in 17 advanced economies since 1870. Since World War II, elevated mortgage credit-to-GDP ratios predict financial crises to the same degree as nonmortgage credit-to-GDP ratios. Furthermore, in predicting recession severity since World War II, the mortgage credit-to-GDP ratio at the beginning of the recession plays an especially important role.

The prominence of household debt is also found in emerging economies. Bahadir and Gumus (2016) focus on Argentina, Brazil, Chile, Korea, Mexico, South Africa, Thailand, and Turkey, and they show that household debt-to-GDP ratios in almost all of these countries have risen substantially since the early 1990s. In contrast, business credit-to-GDP ratios have been relatively stable. They also show significant comovement between household credit and real economic outcomes such as output, consumption, and investment. Increases in household credit are also associated with substantial real exchange rate appreciations. In contrast, changes in business credit have weaker correlations with other real economic outcomes. They use these stylized facts to build a model to distinguish whether shocks to household credit or business credit are driving the real economy. One insight from the model is that household credit shocks are different from business credit shocks in their tendency to simultaneously boost the real exchange rate and increase employment in the nontradable sector.

In Mian, Sufi, and Verner (2017a), we build on this model to show that a credit expansion to businesses that boosts productivity is inconsistent with a simultaneous price increase for nontradable goods and employment growth concentrated in the nontradable sector. In a sample of 36 countries with data back to 1970, we show that household debt booms are associated with a rise in the nontradable to tradable employment ratio, a rise in the nontradable to tradable output ratio, and a rise in the nontradable price to tradable price ratio. In contrast, a rise in firm debt is uncorrelated with these outcomes. This pattern suggests that the household demand channel is dominant.

In addition, in Mian, Sufi, and Verner (2017a), we test these predictions in an evaluation of bank deregulation in the 1980s. As mentioned above, states with a more deregulated banking system as of 1983 experienced a more amplified business cycle from 1983 to 1992. We show that the relative increase in employment in early deregulation states during the expansionary period was concentrated in the nontradable and construction sectors. Furthermore, early deregulation states saw no relative increase in employment in the tradable sector, even among small firms where bank credit is particularly important. The employment patterns are more supportive of credit supply expansion operating through household demand than an expansion in productive capacity by businesses. At the same time, nominal wage growth was substantially stronger in early deregulation states, further supporting the importance of the boost in household demand.

A similar pattern is found among peripheral European countries during the credit expansion period of 2002 to 2007 (Mian, Sufi, and Verner 2017a). Countries in the eurozone with the largest decline in real interest rates experienced employment growth from 2002 to 2007 in the nontradable and construction sectors of 12 to 14 percent, while employment in the tradable sector actually fell 7 percent. Inflation rates were higher in these peripheral countries during this time period, as was nominal wage growth.

Kalantzis (2015) uses a sample of 40 countries from 1970 to 2010. The study isolates 47 episodes of large capital inflows; many are associated with well-known financial or capital account liberalizations such as in Latin America in the 1970s and 1990s, Nordic countries in the 1980s, and Asian countries in the 1990s. He finds that large capital inflows predict a shift of resources from the tradable to nontradable sector. The size of the nontradable sector relative to the tradable sector increases on average by 4 percent relative to normal times.

Explaining the Severity of the Bust

Debt and the Initial Drop in Demand

What makes the recessions that follow expansions in household debt so severe? The initial culprit appears to be a significant drop in household demand. In the Great Recession, for example, in Mian and Sufi (2010), we show that household spending fell substantially even before the heart of the financial crisis in September 2008. In international data, the IMF (2017) study finds a substantial drop in consumption in the aftermath of household debt expansions. Furthermore, both studies find that when a recession does occur, the drop in consumption is stronger in areas where household debt rose the most prior to the recession. Individual-level data also shows that those taking on the most debt during the expansion phase of the credit cycle cut spending the most during the ensuing economic downturn (for evidence from the United Kingdom, see Bunn and Rostom 2015; for Denmark, Andersen, Duus, and Jensen 2014; for a sample of European households, IMF 2017). This channel from high household leverage to a fall in demand was first articulated as the *debt deflation hypothesis* by Irving Fisher (1933), who pointed out that an economic slowdown would raise the real burden of debt, which would further slow the economy through reduced aggregate demand.

Isolating this channel is challenging because other factors that may also interact with economic shocks are often correlated with household leverage. A clear-cut case in favor of Fisher's debt-deflation hypothesis can be found in the Verner and Gyöngyösi (2017) study of Hungary. Some Hungarian households borrowed in Hungarian forint during the 2000s while others borrowed in Swiss francs. This choice of borrowing currency was partly dictated by bank branching networks and was uncorrelated with pre-2008 levels of leverage or growth in house prices, unemployment, or consumption. The sudden appreciation of the Swiss franc in 2008 by over 30 percentage points greatly increased the real burden of debt for a significant fraction of Hungarian households. This sudden rise in the real debt burden generated a sharp decline in household spending.

The drag of debt burdens on consumption during an economic downturn can also be seen in research evaluating a relief in debt payments during the Great Recession in the United States. Di Maggio et al. (2017) exploit variation in the timing of resets on adjustable rate mortgages to show that a 50 percent reduction in mortgage payments boosts spending on autos by 35 percent. They also find that households with low income and low housing wealth see the strongest consumption response to the decline in mortgage payments. In an alternative approach, Agarwal et al. (2017) use regional variation in the implementation of the Home Affordable Modification Program and the Home Affordable Refinancing Program to show that lower mortgage payments associated with the program increased spending. Some of their evidence also suggests that the response was stronger among more indebted borrowers.

Microeconomic studies reveal one reason why the drop in aggregate consumption is so large after debt expansion: debtors have a higher marginal propensity to consume out of wealth and income shocks than those without debt. For example, in Mian, Rao, and Sufi (2013), we show that during the 2006 to 2009 period, households living in zip codes with higher leverage cut back more on spending for the same decline in house prices. Similarly, Baker (forthcoming) shows that Americans with higher debt burdens cut spending by substantially more in response to the same decline in income during the Great Recession. The higher marginal propensity to consume among debtors is an important feature in explaining the severity of recessions following household debt expansions.

Subdued Growth and the Rise in Unemployment

The fact that leveraged households cut spending dramatically after a debt expansion does not, by itself, explain the decline in growth nor the increase in unemployment. For example, the decline in demand by indebted households could trigger a decline in interest rates, thereby boosting demand from less-indebted households or boosting investment by firms. An exchange rate depreciation could increase net exports. However, a variety of frictions prevent such adjustment.

Many countries find themselves at the zero lower bound on nominal interest rates in the aftermath of large expansions in household debt. As illustrated by Hall (2011) and Eggertsson and Krugman (2012), an economy that hits the zero lower bound during the period in which leveraged households cut demand is plagued with a real interest rate that is "too high." As a result, less-leveraged households do not boost spending sufficiently to offset the decline in demand coming from leveraged households. This friction is aggravated by the fact that consumption of less-leveraged households may in general be less-sensitive to credit conditions and interest rates (for example, Sufi 2015; Agarwal, Chomsisengphet, Mahoney, and Stroebel 2018). Households that in normal times have the highest sensitivity of consumption to interest rates and credit availability find themselves either unwilling or unable to borrow in the midst of the downturn that follows credit booms.

Price rigidities play an important role. For example, the negative effect of household debt expansion on subsequent growth is larger in countries with less-flexible exchange rate regimes (Mian, Sufi, and Verner 2017b; IMF 2017). In addition, the effect of a change in household debt on subsequent growth is nonlinear: a large decline in household debt does not predict subsequently stronger growth, but a large increase in household debt predicts subsequently weaker growth (Mian, Sufi, and Verner 2017b). Both of these results suggest that the inability of prices to fall after a debt expansion is one reason the recession is severe.

The aggregate decline in demand quickly spills over into the labor market. Downward nominal wage rigidity is an important reason. For example, Schmitt-Grohé and Uribe (2016) examine the nominal labor cost index for peripheral European countries from 2000 to 2011. Nominal labor costs rose dramatically from 2000 to 2008, but then stayed high from 2008 to 2011 as the unemployment rate jumped from 6 to 14 percent. There is also evidence of significant downward wage rigidity at the state level in the aftermath of the 1980s credit supply expansion in the United States. After the substantial relative nominal wage growth during the credit supply expansion from 1982 to 1989 in early deregulation states, unemployment rose sharply but nominal wages adjusted downward only slowly. Even by 1995, nominal wages remained relatively higher in early deregulation states (Mian, Sufi, and Verner 2017a).

County-level analysis within the United States after the Great Recession also shows the importance of such rigidities. In counties with the largest decline in housing net worth and consumer demand, job losses in the nontradable sector (like retail and restaurant jobs) were severe. However, there was no relative expansion in employment in the tradable sector in these same counties. At least some of the lack of expansion in tradable employment in these counties appears to be related to wage rigidity (Mian and Sufi 2014a).[3] Verner and Gyöngyösi (2017) find similar evidence in Hungary after the depreciation of the local currency in 2008. Areas that experienced a sudden rise in debt burdens see a sharp decline

[3] Beraja, Hurst, and Ospina (2016) show that wages declined more in states where employment fell by the most during the Great Recession, and they argue the data are consistent with only a "modest degree of wage stickiness."

in employment catering to local demand. But wages decline only modestly, and there is no increase in employment among firms operating in the tradable sector.

More generally, recent research suggests that any shock that leads to a large rise in unemployment in the short-term may have large and persistent effects on the labor force and large spillovers onto local economic activity (for example, Acemoglu, Autor, Dorn, Hanson, and Price 2016; Yagan 2017; Acemoglu and Restrepo 2017). If a large drop in household demand generates a substantial rise in unemployment, we should expect the consequences to be large and long-lived.

Foreclosures and a Fall in House Prices

Debt also depresses economic activity during the bust because of forced asset sales. Several studies have investigated how residential foreclosures put downward pressure on house prices and economic activity. In Mian, Sufi, and Trebbi (2015), we exploit variation across states in foreclosure judicial requirements and show that such variation has a strong effect on foreclosure propensity. However, such variation is uncorrelated with the propensity of households to default on their mortgages and uncorrelated with a number of other observable variables. The higher foreclosure propensity in non-judicial-foreclosure states is associated with a decline in house prices, residential investment, and durable goods spending. Using a different identification strategy, Ananbeg and Kung (2014) look at the timing of a listing of a foreclosed property and show that nearby sellers lower their prices in the exact week that the foreclosed property is listed.

In other approaches, Gupta (2016) isolates exogenous variation in foreclosures using shocks to interest rates resulting from details in adjustable rate mortgage contracts. He finds that a foreclosure leads to further foreclosures and lower house prices in the surrounding area. Furthermore, a foreclosure leads to difficulty in refinancing mortgages into lower rates for those living close to the foreclosed property, as banks tend to use the depressed foreclosure price as a comparison. Using a quantitative model of the housing market, Guren and McQuade (2015) find that foreclosures during the Great Recession exacerbate US aggregate house price declines by 62 percent and nonforeclosure price declines by 28 percent. Verner and Gyöngyösi (2017) find similar effects in Hungary.

Banking Crises

Another reason for the severity of the recessions following an expansion in credit supply is that the resulting crunch can involve a severe tightening of credit supply that may affect all households and businesses.

Households in the United States living in zip codes with high leverage and a decline in house prices during the Great Recession faced a particularly acute contraction in credit supply. Home equity limits and credit card limits fell significantly more in these zip codes relative to the rest of the country (Mian, Rao, and Sufi 2013). First-time home-buying contracted more severely for low credit score versus high credit score individuals, which also suggests a tightening of credit supply (Bhutta 2015).

In addition, the US banking crisis in the Great Recession led to a decline in employment and consumption that spread beyond leveraged households. Firms borrowing from banks that were most exposed to the banking crisis witnessed a larger decline in employment during the Great Recession (Chodorow-Reich 2014). Employment losses in the nontradable sector were particularly large in counties with a large drop in demand, and these employment losses were concentrated among firms with weak balance sheets that were likely most exposed to the adverse credit conditions during the Great Recession (Giroud and Mueller 2017). On the spending side, the collapse in the asset-backed commercial paper market led to a collapse in the availability of nonbank auto loan financing. As a result, counties that traditionally relied on nonbank auto loan financing witnessed a substantial decline in auto purchases (Benmelech, Meisenzahl, and Ramcharan 2017).

A banking crisis disrupts economic activity for a variety of reasons, in line with the financial accelerator view of Bernanke (1983), Bernanke and Gertler (1989), and Kiyotaki and Moore (1997). However, banking crises should not be viewed independently from the expansion in household debt that often precedes them. After all, household debt is a key asset held by banks, and so a rise in household defaults will directly affect the banking sector. As mentioned above, Jordà, Schularick, and Taylor (2016) show that a rise in mortgage credit-to-GDP ratios predicts banking crises. Additionally, they show that recessions associated with high mortgage debt growth *and* a banking crisis are the most severe.

Longer-term Distortions

A credit boom distorts the economy, which can then make the subsequent recession more severe and protracted. One such distortion is the large increase in employment in the retail and construction sectors. Areas of the United States with substantial housing booms experienced substantial improvement in labor market opportunities for young men and women. As a result, these areas witnessed lower college enrollment, especially at two-year colleges. After the bust, many of these individuals did not return to college "suggesting that reduced educational attainment is an enduring effect of the recent housing cycle" (Charles, Hurst, and Notowidigdo forthcoming).

In another study of across-sector labor reallocation during periods of rapid private credit growth, Borio, Kharroubi, Upper, and Zampolli (2016) find that workers systematically moved into low-productivity growth sectors, which in turn led to lower productivity growth after the recession. This pattern was especially prevalent in recessions associated with financial crises. Gopinath, Kalemli-Ozcan, Karabarbounis, and Villegas-Sanchez (2015) show how credit supply expansion lowered productivity growth among Spanish manufacturing firms between 1999 and 2012 by directing funds toward higher net worth firms that were not necessarily more productive.

Theoretical Foundations

What existing models help us to understand the credit-driven household demand channel? In this section, we first discuss existing theoretical research that treats credit supply expansion as exogenous, and then we turn to theoretical models that can explain how credit supply expansion leads to predictable boom-bust cycles.

Credit Supply Expansion as an Exogenous Shock

Much of the existing theoretical research treats credit supply expansion as an exogenous shock. As one example, Schmitt-Grohé and Uribe (2016) examine a small open economy with a pegged exchange rate. In one exercise, they assume an exogenous decline in the country interest rate, which subsequently reverses. As another example, in the model of Justiniano et al. (2015), total lending by savers is limited exogenously, and a credit supply expansion in their model is a relaxation of this lending constraint.

Other studies have modeled a credit shock as a relaxation of loan-to-value or payment-to-income constraints. While these are components of debt booms, there are drawbacks in treating them as the main force driving credit supply expansions. As Justiniano, Primiceri, and Tambalotti (2015) point out, a relaxation of a loan-to-value constraint by itself actually leads to an increase in mortgage interest rates, which is counterfactual for most episodes. Kiyotaki, Michaelides, and Nikolov (2011) and Kaplan et al. (2017) argue that a relaxation of loan-to-value constraints alone cannot explain the rise in house prices that is typical of these credit booms.

As a result of these issues, models that rely on relaxation of these loan constraints typically also contain a second force that is necessary to fit the facts. Favilukis, Ludvigson, and Van Nieuwerburgh (2017) consider both financial market liberalization, which consists of a loosening of a loan-to-value constraint on mortgages and lower transactions costs associated with obtaining a mortgage, along with an influx of foreign funds into the domestic risk-free bond market. The combination of these shocks is necessary to generate an increase in household debt, an increase in house prices, and a steady or declining risk-free interest rate. Similarly, Greenwald (2016) models a credit supply expansion as a simultaneous loosening of a payment-to-income constraint on mortgages and a decline in the real interest rate. Again, both forces are necessary to generate the observed patterns in housing markets during the 2000 to 2007 period in the United States. Garriga, Manuelli, and Peralta-Alva (2018) build a model where there are exogenous changes in both loan-to-value ratios and mortgage interest rates. They conclude that a decline in mortgage interest rates is the more important quantitative force leading to house price appreciation, but that the interaction of the two forces can amplify the effect of mortgage rates on home values.

Another important point is that credit supply expansions manifest themselves far beyond a higher allowed loan-to-value or price-to-income ratio. We concur with Favilukis et al. (2012) who write that "the behavior of combined loan-to-value ratios in the boom and bust does not do full justice to several aspects of increased availability of mortgage credit." As they point out, the 2000 to 2007 mortgage credit expansion in the United States was associated with previously rationed borrowers receiving credit, new mortgage contracts, and reduced asset and income verification by lenders. A narrow focus on loan-to-value and payment-to-income ratios misses many dimensions of credit supply expansion episodes.

Rational Expectations and Credit-Driven Externalities

What models can help to explain the predictable boom-bust episode gnerated by an expansion in credit supply? One class of models relies on credit-driven externalities. A temporary positive shock to credit supply occurs, but all households share a common understanding that the shock will disappear at some time in the future. However, despite rational expectations and the transient nature of credit expansion, there is "overborrowing" from a social planner's perspective, and such overborrowing generates a boom-bust cycle in both credit and the real economy.

One such reason for overborrowing is the presence of aggregate demand externalities (for example, Eggertsson and Krugman 2012; Farhi and Werning 2016; Ríos-Rull and Huo 2016; Korinek and Simsek 2016; Schmitt-Grohé and Uribe 2016; Guerrieri and Lorenzoni 2017). In these models, there is a friction such as nominal wage rigidity or a lower bound on the real interest rate that prevents the economy from adjusting when credit contracts and there is a drop in demand from leveraged households. Households do not internalize the effect of their future decline in demand on the income of other households, and they therefore rationally take on more debt than is socially optimal.

Another reason for overborrowing is the presence of pecuniary externalities due to "fire sales" as discussed in Shleifer and Vishny (1992), Kiyotaki and Moore (1997), Caballero and Krishnamurthy (2001), Lorenzoni (2008), Bianchi (2011), Dávila (2015), and others. Suppose that an asset, such as a house, is used as collateral for borrowing. If households borrow in the present, they will tend to drive up the price of the asset. After a negative shock, households will be forced to de-lever by fire-selling the collateral, which reduces the price of collateral and hence tightens the borrowing constraint. In this way, the collateral price channel adds to the aggregate demand externality. In both cases, households may rationally decide to take on more debt during an expansion than is socially optimal because they do not internalize the effect of their actions on others during the credit contraction.

Heterogeneous Beliefs and Behavioral Biases

The rational expectations framework with a temporary, self-reverting credit shock can offer an explanation for why an expansion in credit supply leads to a boom-bust cycle. However, an explanation based on rational expectations and externalities has one major problem: it predicts that individuals during a credit boom anticipate a slowdown in the economy. This prediction is counterfactual. As noted earlier, economic forecasters systematically over-predict future GDP growth during credit booms. In addition, market participants often fail to foresee the correction in asset prices that typically occurs in the aftermath of credit booms. For example, high levels of bank credit also seem to be associated with a predictable crash in equity prices for banks (Baron and Xiong 2017), and banks that expand credit most rapidly have predictably worse returns in the subsequent years (Fahlenbrach, Prilmeier, and Stulz 2017). For these reasons, the rational expectations model with common beliefs is unlikely to explain the predictable boom-bust cycles we witness in the data.

One alternative is to move away from the assumption of common beliefs. Geanakoplos (2010) builds a theory of endogenous leverage cycles in which households differ in their level of optimism about the economy. Burnside, Eichenbaum, and Rebelo (2016) also build a model in which belief heterogeneity plays an important role in explaining boom-bust cycles in the housing market. Greater availability of credit in such an environment enables optimists to increase leverage, to buy more of the collateralized asset, and therefore to raise asset prices. A positive credit supply shock results in giving the optimists' expectations greater weight in market prices. As a result, credit, asset prices, and market expectations rise collectively.

However, even a small negative shock bankrupts the optimists who are highly leveraged because of their exuberant beliefs. Consequently, these optimists must dump assets in the market, and the only households with positive net worth who can buy these assets are the pessimists. Asset prices fall, which further reinforces the original wave of fire sales and credit contraction. This endogenous boom-bust leverage cycle may interact with frictions in the macroeconomy discussed earlier, thereby generating a boom-bust cycle in the real economy.

Another approach, relying on behavioral biases, has been emphasized at least since Minsky (2008) and Kindleberger (1978). This approach is consistent with empirically observed errors in expectations and can also generate credit cycles. For example, in Gennaioli, Shleifer, and Vishny (2012), investors neglect tail risks, which leads to aggressive lending by the financial sector via debt contracts. In Landvoigt (2016), the lending boom is instigated when creditors underestimate the true default risk of mortgages. In Greenwood, Hanson, and Jin (2016), exuberant credit market sentiment boosts lending because lenders mistakenly extrapolate previously low defaults when granting new loans. Bordalo, Gennaioli, and Shleifer (2017) provide micro-foundations for such mistakes by lenders, which they refer to as "diagnostic expectations."

These behavioral biases can be viewed as part of a process that leads to credit supply expansions. For example, perhaps lenders begin lending to lower-credit-quality borrowers because they mistakenly believe that the probability of default for such borrowers is lower than it is. Or perhaps mortgage credit spreads fall because lenders become more optimistic about house price growth, as in Kaplan, Mitman, and Violante (2017).

A further advantage of the behavioral models is that they may be able to generate endogenously a reversal in credit supply after an expansion driven by behavioral biases. For example, Bordalo, Gennaioli, and Shleifer (2017) generate predictable reversals in credit supply given the biased expectations formed by investors. As they note, "following this period of narrow credit spreads, these spreads predictably rise on average ... while investment and output decline ..." While the exact timing of the reversal is not known, a rise in credit supply driven by lender optimism eventually reverts as lenders become pessimistic.

What Drives Credit Supply Expansion

Much of the work on the credit-driven household demand channel takes the credit supply expansion as given. But what kind of shock leads to an expansion in credit supply? We should admit that we have now entered a more speculative part of this essay. The evidence currently available is less conclusive on this question.

In our view, the most likely initial shock is one that creates an excess of savings relative to investment demand in some part of the global financial system, what we call a "financial excess." This initial shock can be amplified by behavioral biases, financial innovation, and even by malfeasance within the financial sector.

Perhaps the most popular version of such a financial excess is the "global savings glut" hypothesis articulated in Bernanke (2005), which focuses on the "metamorphosis of the developing world from a net user to a net supplier of funds to international capital markets." In response to financial crises in the late 1990s and early 2000s, governments in emerging markets began to accumulate foreign reserves, typically in the form of US-dollar denominated assets. In turn, this shift led to declining global interest rates, the rise of dollar-denominated assets, and current account deficits in many advanced economies. Alpert (2013) and Wolf (2014) both place high importance on the global savings glut as a reason for the boom and bust in economic activity from 2000 to 2010 in many advanced economies.

The combination of OPEC price increases in the 1970s and the Latin American debt crisis of the early 1980s offers another example. Pettis (2017) points to financial excesses created by OPEC countries: "[I]n the early 1970s, for example, as a newly assertive OPEC drove up oil prices and deposited their massive surplus earnings in international banks, these banks were forced to find borrowers to whom they could recycle these flows. They turned to a group of middle-income developing countries, including much of Latin America." Devlin (1989) also points to the dramatic increase in oil prices in 1973 and 1974 as a source of credit supply expansion. As he points out, a large fraction of the surplus dollars earned by oil-producing countries entered the international private banking system. In response, "banks become much more active lenders, and the scope of their operations expanded enormously." Similarly, Folkerts-Landau (1985) writes that "the international payments imbalances generated by the oil price increase of 1973 provided an unprecedented opportunity for the international credit markets to expand."

External debt of non-oil developing countries increased from $97 billion in 1973 to $505 billion in 1982 (Bernal 1982). During this credit expansion, syndicated bank loan interest spreads over LIBOR on loans to these countries fell from 1.6 to 0.7 percent. Similarly, Devlin (1989) writes: "By 1977 not only did loan volume [to Latin America] continue to rise but the terms of lending softened as the situation moved back into a so-called borrowers' market. ... [B]eginning in 1977 spreads came down sharply and maturities were commonly awarded in excess of five years. The trend toward lower spreads and longer maturities became sharply accentuated in 1978 to 1980."

In both of these examples, a set of countries experienced an expansion in credit supply because of financial excesses created in international markets. Examples of a shock leading to financial excesses in a closed-economy setting are also available, if less common.

One example proposed by Kumhof, Rancière, and Winant (2015) is the rise in income inequality. They look at rising inequality prior to both the Great Depression and Great Recession. In both episodes, there was a simultaneous large increase in debt-to-income ratios among lower- and middle-income households. In their model, a rise in income inequality leads to more funds entering the financial system as high-income households have a preference for wealth accumulation and therefore a high marginal propensity to save. Thus a rise in income inequality acts as a credit supply expansion to middle- and lower-income households. The model also predicts a decline in the interest rate on household borrowing, which is consistent with the empirical evidence.

Other possible domestic sources of credit supply expansions include financial liberalization and financial deregulation, especially for smaller open economies. For example, Kindleberger and Aliber (2005) write that "a particular recent form of displacement that shocks the system has been financial liberalization or deregulation in Japan, the Scandinavian countries, some of the Asian countries, Mexico, and Russia. Deregulation has led to monetary expansion, foreign borrowing, and speculative investment." Two studies mentioned above exploit variation across the United States in banking deregulation: Di Maggio and Kermani (2017) and Mian, Sufi, and Verner (2017a). Both show that states that experience more deregulation see a bigger increase in credit supply during aggregate credit expansion episodes.

The Latin American debt crisis of the early 1980s was also preceded by a round of deregulation that scholars have pointed to as a source of the rapid expansion in debt (for example, Diaz-Alejandro 1985). As McKinnon (1984) notes, "[T]he case of the Southern Cone in the 1970s and early 1980s is hardly very pure; in this period virtually all less-developed countries overborrowed, and then got themselves into a debt crisis. This era was complicated by a recycling from the oil shock on the one hand and then what I consider to be a major breakdown in the public regulation of risk-taking Western banks on the other. The result was gross overlending by banks in the world economy at large and to the Third World in particular."

The Scandinavian banking crises of the late 1980s and early 1990s also followed a financial deregulation. In his overview of the banking crises in Norway, Finland, and Sweden, Englund (1999) concludes that "newly deregulated credit markets after 1985 stimulated a competitive process between financial institutions where expansion was given priority." Jonung, Kiander, and Vartia (2008) focus on the banking crises in Sweden and Finland. They write, "the boom-bust process starts with a deregulation of financial markets leading to a rapid inflow of capital to finance domestic investments and consumption."

From the perspective of a given country or state, deregulation of the financial sector may lead to capital inflows and a credit supply expansion. In this sense, deregulation is the shock that leads to an expansion in credit supply from the perspective of the country or state. This narrative tells us where credit lands, but it still leaves open the question of why so much credit is looking for a place to land in the first place. For this reason, we give more importance to the view that financial excess is the initial shock starting the expansion process. But the level of regulation or efforts at deregulation will help determine where credit lands during credit supply expansions.

Directions for Future Research

The credit-driven household demand channel is the idea that credit supply expansions operating through household demand are an important source of business cycles. The Great Recession is the most prominent example, but this phenomenon is present in many episodes the world has witnessed over the past 50 years.

In this article, we have presented evidence supporting the three main pillars of the credit-driven household demand channel. First, credit supply expansions lead to a boom-bust cycle in household debt and real economic activity. Second, expansions tend to affect the real economy through a boost to household demand as opposed to an increase in productive capacity of firms. Third, the downturn is driven initially by a decline in aggregate demand which is further amplified by nominal rigidities, constraints on monetary policy, banking sector disruptions, and legacy distortions from the boom.

The credit-driven household demand channel is distinct from traditional financial accelerator models (Bernanke and Gertler 1989; Kiyotaki and Moore 1997; Bernanke, Gertler, and Gilchrist 1999), primarily due to the centrality of households as opposed to firms in explaining the real effects of credit supply expansions. In addition, while there are examples of financial accelerator models that focus on the expansion phase of the credit cycle and explore the importance of behavioral biases (Bernanke and Gertler 2000), these factors play a more central role in the credit-driven household demand channel.

There remain a number of open questions related to the credit-driven household demand channel. For example, what is the fundamental source that causes lenders to increase credit availability? Why do some credit booms end in a crash while others may not (for example, Gorton and Ordonez 2016)? What is the sequence of events that initiates the crisis stage?

The policy implications of this idea need more exploration, too. Should regulators impose macroprudential limits on household debt? Should monetary policymakers "lean against the wind" during credit supply expansions? Should the government encourage the use of debt contracts? During the bust, what is the most effective policy at limiting the damage coming from the collapse in aggregate demand? We have offered preliminary answers to these questions elsewhere (Mian and Sufi 2014b, 2017c), but definitive answers require more investigation on both the theoretical and empirical fronts.

Finally, while we have emphasized the business cycle implications of the credit-driven household demand channel, the analysis presented here may prove relevant for longer-run growth considerations. Since 1980, advanced economies of the world have experienced four key trends: 1) Most advanced economies have seen a substantial rise in wealth and income inequality. 2) Borrowing costs have fallen dramatically, especially on risk-free debt. 3) Household debt-to-GDP ratios have increased substantially, and most of bank lending is now done via mortgages (Jordà, Schularick, and Taylor 2016). 4) Finally, the financial sector has grown as a fraction of GDP. Are these four patterns linked? Can they help explain why global growth for advanced economies has been so weak since the onset of the Great Recession in 2007 (for example, Summers 2014)? One preliminary idea is that there is a global excess supply of savings coming from both the rise in income inequality in advanced economies and the tendency of some emerging economies to export capital to advanced economies. This excess savings leads to growth in the financial sector, a decline in interest rates, and a rise in household debt burdens of households in advanced economies outside the very top of the income distribution. But at this stage, the connection of these patterns to growth remains a more open question.

■ *We thank Daron Acemoglu, Mark Gertler, Gordon Hanson, Andrei Shleifer, Jeremy Stein, Timothy Taylor, Emil Verner, and Robert Vishny for comments that substantially improved the draft. We thank Tu Cao and Michael Varley for excellent research assistance. We also thank the Julis Rabinowitz Center for Public Policy and Finance at Princeton and the Initiative on Global Markets and the Fama-Miller Center at Chicago Booth for financial support.*

References

Acemoglu, Daron, David Autor, David Dorn, Gordon H. Hanson, and Brendan Price. 2016. "Import Competition and the Great US Employment Sag of the 2000s." *Journal of Labor Economics* 34(S1, Part 2): 141–98.

Acemoglu, Daron, and Pascual Restrepo. 2017. "Robots and Jobs: Evidence from US labor Markets." NBER Working Paper 23285.

Adelino, Manuel, Antoinette Schoar, and Felipe Severino. 2014. "Credit Supply and House Prices: Evidence from Mortgage Market Segmentation." Available at SSRN: https://ssrn.com/abstract=1787252.

Adelino, Manuel, Antoinette Schoar, and Felipe Severino. 2017. "Dynamics of Housing Debt in the Recent Boom and Bust." *NBER Macroeconomics*

Annual 32: 261–311.

Agarwal, Sumit, Gene Amromin, Itzhak Ben-David, Souphala Chomsisengphet, Tomasz Piskorski, and Amit Seru. 2017. "Policy Intervention in Debt Renegotiation: Evidence from the Home Affordable Modification Program." *Journal of Political Economy* 125(3): 654–712.

Agarwal, Sumit, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel. 2018. "Do Banks Pass through Credit Expansions to Consumers Who Want to Borrow? Evidence from Credit Cards." *Quarterly Journal of Economics* 133(1): 129–90.

Alpert, Daniel. 2013. *The Age of Oversupply: Overcoming the Greatest Challenge to the Global Economy.* Portfolio.

Andersen, Asger Lau, Charlotte Duus, and Thais Lærkholm Jensen. 2014. "Household Debt and Consumption During the Financial Crisis: Evidence from Danish Micro Data." Working Paper 89, Danmarks Nationalbank.

Anenberg, Elliot, Aurel Hizmo, Edward Kung, and Raven Molloy. 2017. "Measuring Mortgage Credit Availability: A Frontier Estimation Approach." FEDS Working Paper 2017-101, Financial and Economics Discussion Series, Board of Governors of the Federal Reserve System.

Anenberg, Elliot, and Edward Kung. 2014. "Estimates of the Size and Source of Price Declines Due to Nearby Foreclosures." *American Economic Review* 104(8): 2527–51.

Bahadir, Berrak, and Inci Gumus. 2016. "Credit Decomposition and Business Cycles in Emerging Market Economies." *Journal of International Economics* 103: 250–62.

Baker, Scott R. Forthcoming. "Debt and the Response to Household Income Shocks: Validation and Application of Linked Financial Account Data." *Journal of Political Economy***.**

Baron, Matthew, and Wei Xiong. 2017. "Credit Expansion and Neglected Crash Risk." *Quarterly Journal of Economics* 132(2): 713–64.

Benmelech, Efraim, Ralf R. Meisenzahl, and Rodney Ramcharan. 2017. "The Real Effects of Liquidity During the Financial Crisis: Evidence from Automobiles." *Quarterly Journal of Economics* 132(1): 317–65.

Beraja, Martin, Erik Hurst, and Juan Ospina. 2016. "The Aggregate Implications of Regional Business Cycles." March 15. http://faculty.chicagobooth.edu/erik.hurst/research/regional_paper_submit.pdf.

Bernal, Richard. 1982. "Transnational Banks, the International Monetary Fund and External Debt of Developing Countries." *Social and Economic Studies* 31(4): 71–101.

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

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

Bernanke, Ben, and Mark Gertler. 1989. "Agency Costs, Net Worth, and Business Fluctuations." *American Economic Review* 79(1): 14–31.

Bernanke, Ben, and Mark Gertler. 2000. "Monetary Policy and Asset Price Volatility." NBER Working Paper 7559.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." Chap. 21 in *Handbook of Macroeconomics* vol. 1, edited by John Taylor and Michael Woodford. Elsevier.

Bhutta, Neil. 2015. "The Ins and Outs of Mortgage Debt during the Housing Boom and Bust." *Journal of Monetary Economics* 76: 284–98.

Bianchi, Javier. 2011. "Overborrowing and Systemic Externalities in the Business Cycle." *American Economic Review* 101(7): 3400–26.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2017. "Diagnostic Expectations and Credit Cycles." *Journal of Finance* 73(1).

Borio, Claudio. 2014. "The Financial Cycle and Macroeconomics: What Have We Learnt?" *Journal of Banking and Finance* 45: 182–198.

Borio, Claudio, Enisse Kharroubi, Christian Upper, and Fabrizio Zampolli. 2016. "Labour Reallocation and Productivity Dynamics: Financial Causes, Real Consequences." BIS Working Paper 534.

Bunn, Philip, and May Rostom. 2015. "Household Debt and Spending in the United Kingdom." Bank of England Working Paper 554.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo. 2016. "Understanding Booms and Busts in Housing Markets." *Journal of Political Economy* 124(4): 1088–1147.

Caballero, Ricardo J., and Arvind Krishnamurthy. 2001. "International and Domestic Collateral Constraints in a Model of Emerging Market Crises." *Journal of Monetary Economics* 48(3): 513–548.

Charles, Kerwin Kofi, Erik Hurst, and Matthew J. Notowidigdo. Forthcoming. "Housing Booms and Busts, Labor Market Opportunities, and College Attendance." *American Economic Review.*

Chodorow-Reich, Gabriel. 2014. "The Employment Effects of Credit Market Disruptions: Firm-level Evidence from the 2008–9 Financial Crisis." *Quarterly Journal of Economics* 129(1): 1–59.

Dávila, Eduardo. 2015. "Dissecting Fire Sales Externalities." 2015. Unpublished paper.

Demyanyk, Yuliya, and Otto Van Hemert. 2011. "Understanding the Subprime Mortgage Crisis." *Review of Financial Studies* 24(6): 1848–80.

Devlin, Robert. 1989. *Debt and Crisis in Latina America: The Supply Side of the Story*. Princeton University Press.

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

Di Maggio, Marco, and Amir Kermani. 2017. "Credit-Induced Boom and Bust." *Review of Financial Studies* 30(11): 3711–58.

Di Maggio, Marco, Amir Kermani, Benjamin J. Keys, Tomasz Piskorski, Ramcharan Rodney, Amit Seru, and Vincent Yao. 2017. "Interest Rate Pass-Through: Mortgage Rates, Household Consumption, and Voluntary Deleveraging." *American Economic Review* 107(11): 3550–88.

Drehmann, Mathias, Mikael Juselius, and Anton Korinek. 2017. "Accounting for Debt Service: The Painful Legacy of Credit Booms." BIS Working Papers 645, June.

Eggertsson, Gauti B., and Paul Krugman. 2012. "Debt, Deleveraging, and the Liquidity Trap: A Fisher-Minsky-Koo Approach." *Quarterly Journal of Economics* 127(3): 1469–1513.

Englund, Peter. 1999. "The Swedish Banking Crisis: Roots and Consequences." *Oxford Review of Economic Policy* 15(3): 80–97.

Fahlenbrach, Rüdiger, Robert Prilmeier, René M. Stulz. 2017. "Why Does Fast Loan Growth Predict Poor Performance for Banks?" *Review of Financial Studies* 31(3): 1014–63.

Farhi, Emmanuel, and Iván Werning. 2016. "A Theory of Macroprudential Policies in the Presence of Nominal Rigidities." *Econometrica* 84(5): 1645–1704.

Favara, Giovanni, and Jean Imbs. 2015. "Credit Supply and the Price of Housing." *American Economic Review* 105(3): 958–92.

Favilukis, Jack, David Kohn, Sydney C. Ludvigson, and Stijn Van Nieuwerburgh. 2012. "International Capital Flows and House Prices: Theory and Evidence." Chap. 6 in *Housing and the Financial Crisis*, edited by Edward L. Glaeser and Todd Sinai. University of Chicago Press.

Favilukis, Jack, Sydney C. Ludvigson, and Stijn Van Nieuwerburgh. 2017. "The Macroeconomic Effects of Housing Wealth, Housing Finance, and Limited Risk Sharing in General Equilibrium." *Journal of Political Economy* 125(1): 140–223.

Fisher, Irving. 1933. "The Debt-Deflation Theory of Great Depressions." *Econometrica* 1(4): 337–57.

Folkerts-Landau, David. 1985. "The Changing Role of International Bank Lending in Development Finance." *IMF Staff Papers* 32(2): 317–63.

Foote, Christopher L., Kristopher S. Gerardi, and Paul S. Willen. 2012. "Why Did So Many People Make So Many Ex Post Bad Decisions? The Causes of the Foreclosure Crisis." NBER Working Paper 18082.

Garriga, Carlos, Rodolfo Manuelli, and Adrian Peralta-Alva. 2018. "A Macroeconomic Model of Price Swings in the Housing Market." Federal Reserve Bank of Saint Louis Working Paper.

Geanakoplos, John. 2010. "The Leverage Cycle." *NBER Macroeconomics Annual 2009*, vol. 24, edited by Daron Acemoglu, Kenneth Rogoff, and Michael Woodford, pp. 1–65. University of Chicago Press.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. 2012. "Neglected Risks, Financial Innovation, and Financial Fragility." *Journal of Financial Economics* 104(3): 452–68.

Giroud, Xavier, and Holger Mueller. 2017. "Firm Leverage, Consumer Demand, and Unemployment during the Great Recession." *Quarterly Journal of Economics* 132(1): 271–316.

Glick, Reuven, and Kevin J. Lansing. 2010. "Global Household Leverage, House Prices, and Consumption." FRBSF Economic Letter 2010-01, Federal Reserve Bank of San Francisco. January 11.

Gopinath, Gita, Sebnem Kalemli-Ozcan, Loukas Karabarbounis, and Carolina Villegas-Sanchez. 2015. "Capital Allocation and Productivity in South Europe." NBER Working Paper 21453, August.

Gorton, Gary, and Guillermo Ordonez. 2016. "Good Booms, Bad Booms." NBER Working Paper 22008.

Greenwald, Daniel L. 2016. "The Mortgage Credit Channel of Macroeconomic Transmission." MIT Sloan Research Paper no. 5184-16. Available at SSRN https://ssrn.com/abstract=2735491.

Greenwood, Robin, Samuel G. Hanson, and Lawrence J. Jin. 2016. "A Model of Credit Market Sentiment." HBS Working Paper no. 17-014, Harvard Business School.

Griffin, John M., and Gonzalo Maturana. 2016a. "Did Dubious Mortgage Origination Practices Distort House Prices?" *Review of Financial Studies* 29(7): 1671–1708.

Griffin, John M., and Gonzalo Maturana. 2016b. "Who Facilitated Misreporting in Securitized Loans?" *Review of Financial Studies* 29(2): 384–419.

Guerrieri, Veronica, and Guido Lorenzoni. 2017b. "Credit Crises, Precautionary Savings, and the Liquidity Trap." *Quarterly Journal of Economics* 132(3): 1427–67.

Gupta, Arpit. 2016. "Foreclosure Contagion and the Neighborhood Spillover Effects of Mortgage Defaults." Unpublished paper.

Hall, Robert E. 2011. "The Long Slump." *American Economic Review* 101(2): 431–69.

International Monetary Fund (IMF). 2012. "Dealing with Household Debt." Chap. 3 in *World Economic Outlook, April 2012: Growth Resuming, Dangers Remain.* IMF.

International Monetary Fund (IMF). 2017. "Household Debt and Financial Stability." Chap. 2 in *Global Financial Stability Report, October 2017: Is Growth at Risk.* IMF.

Jayaratne, Jith, and Philip E. Strahan. 1996. "The Finance–Growth Nexus: Evidence from Bank Branch Deregulation." *Quarterly Journal of Economics* 111(3): 639–670.

Jonung, Lars, Jaakko Kiander, and Pentti Vartia. 2008. "The Great Financial Crisis in Finland and Sweden: The Dynamics of Boom, Bust, and Recovery, 1985–2000." Economic Papers, 350, Economic and Financial Affairs, European Commission. http://ec.europa.eu/economy_finance/publications/pages/publication13551_en.pdf.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2015. "Betting the House." *Journal of International Economics* 96(S2–S18): 29.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2016. "The Great Mortgaging: Housing Finance, Crises and Business Cycles." *Economic Policy* 31(85): 107–152.

Justiniano, Alejandro, Giorgio E. Primiceri, and Andrea Tambalotti. 2015. "Credit Supply and the Housing Boom." NBER Working Paper 20874, January.

Justiniano, Alejandro, Giorgio E. Primiceri, and Andrea Tambalotti. 2017. "The Mortgage Rate Conundrum." NBER Working Paper 23784.

Kalantzis, Yannick. 2015. "Financial Fragility in Small Open Economies: Firm Balance Sheets and the Sectoral Structure." *Review of Economic Studies* 82(3): 1194–1222.

Kaplan, Greg, Kurt Mitman, and Giovanni L. Violante. 2017. "The Housing Boom and Bust: Model Meets Evidence." NBER Working Paper 23694.

Keys, Benjamin J., Tanmoy Mukherjee, Amit Seru, and Vikrant Vig. 2010. "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *Quarterly Journal of Economics* 125(1): 307–362.

Kindleberger, Charles. 1978. *Manias, Panics and Crashes: A History of Financial Crises.* New York; Basic Books.

Kindelberger, Charles P., and Robert Z. Aliber. 2005. *Manias, Panics and Crashes: A History of Financial Crises.* Palgrave Macmillan.

Kiyotaki, Nobuhiro, Alexander Michaelides, and Kalin Nikolov. 2011. "Winners and Losers in Housing Markets." *Journal of Money, Credit and Banking* 43(2–3): 255–96.

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

Korinek, Anton, and Alp Simsek. 2016. "Liquidity Trap and Excessive Leverage." *American Economic Review* 106(3): 699–738.

Krishnamurthy, Arvind, and Tyler Muir. 2017. "How Credit Cycles across a Financial Crisis." NBER Working Paper 23850.

Kroszner, Randall S., and Philip E. Strahan. 2014. "Regulation and Deregulation of the U.S. Banking Industry: Causes, Consequences, and Implications for the Future." Chap. 8 in *Economic Regulation and It's Reform: What Have We Learned?* University of Chicago Press.

Kumhof, Michael, Romain Rancière, and Pablo Winant. 2015. "Inequality, Leverage, and Crises." *American Economic Review* 105(3): 1217–45.

Laibson, David, and Johanna Mollerstrom. 2010. "Capital Flows, Consumption Booms and Asset Bubbles: A Behavioural Alternative to the Savings Glut Hypothesis." *Economic Journal* 120(544): 354–74.

Landvoigt, Tim. 2016. "Financial Intermediation, Credit Risk, and Credit Supply during the Housing Boom." Available at SSRN: https://ssrn.com/abstract=2834074.

Levitin, Adam J., and Susan M. Wachter. 2012. "Explaining the Housing Bubble." *Georgetown Law Journal* 100(4): 1177–1258.

López-Salido, David, Jeremy C. Stein, Egon Zakrajšek. 2017. "Credit-Market Sentiment and the Business Cycle." *Quarterly Journal of Economics* 132(3): 1373–1426.

Lorenzoni, Guido. 2008. "Inefficient Credit Booms." *Review of Economic Studies* 75(3): 809–33.

Martin, Philippe, and Thomas Philippon. 2017. "Inspecting the Mechanism: Leverage and the Great Recession in the Eurozone." *American Economic Review* 107(7): 1904–37.

Mayer, Christopher. 2011. "Housing Bubbles: A Survey." *Annual Review of Economics* 3: 559–77.

McQuade, Timothy, and Adam Guren. 2015. "How Do Foreclosures Exacerbate Housing Downturns?" 2015 Meeting papers, no. 40, Society for Economic Dynamics.

Mertens, Karel, and Morten O. Ravn. 2013. "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States." *American Economic Review* 103(4): 1212–47.

Mian, Atif, Kamalesh Rao, and Amir Sufi. 2013. "Household Balance Sheets, Consumption, and the Economic Slump." *Quarterly Journal of Economics* 128(4): 1687–1726.

Mian, Atif, and Amir Sufi. 2009. "The Consequences of Mortgage Credit Expansion: Evidence from the U.S. Mortgage Default Crisis."

Quarterly Journal of Economics 124(4): 1449–96.

Mian, Atif, and Amir Sufi. 2010. "Household Leverage and the Recession of 2007-09." *IMF Economic Review* 58(1): 74–117.

Mian, Atif, and Amir Sufi. 2011. "House Prices, Home Equity-Based Borrowing, and the US Household Leverage Crisis." *American Economic Review* 101(5): 2132–56.

Mian, Atif, and Amir Sufi. 2014a. "What Explains the 2007–2009 Drop in Employment?" *Econometrica* 82(6): 2197–2223.

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

Mian, Atif, and Amir Sufi. 2017a. "Fraudulent Income Overstatement on Mortgage Applications during the Credit Expansion of 2002 to 2005." *Review of Financial Studies* 30(6): 1832–64.

Mian, Atif R., and Amir Sufi. 2017b. "Household Debt and Defaults from 2000 to 2010: The Credit Supply View." In *Evidence and Innovation in Housing Law and Policy*, edited by L. Fennell and B. Keys, 257–88. Cambridge University Press.

Mian, Atif R., and Amir Sufi. 2017c. "The Macroeconomic Advantages of Softening Debt Contracts." *Harvard Law and Policy Review* 11(1).

Mian, Atif, Amir Sufi, and Francesco Trebbi. 2015. "Foreclosures, House Prices, and the Real Economy." *Journal of Finance* 70(6): 2587–2634.

Mian, Atif R., Amir Sufi, and Emil Verner. 2017a. "How Do Credit Supply Shocks Affect the Real Economy? Evidence from the United States in the 1980s." Available at SSRN: https://ssrn.com/abstract=2971086.

Mian, Atif R., Amir Sufi, and Emil Verner. 2017b. "Household Debt and Business Cycles Worldwide." *Quarterly Journal of Economics* 132(4): 1755–1817.

Minsky, Hyman P. 2008. *Stabilizing an Unstable Economy*, vol. 1. New York: McGraw-Hill.

Pennington-Cross, Anthony, and Souphala Chomsisengphet. 2007. "Subprime Refinancing: Equity Extraction and Mortgage Termination." *Real Estate Economics* 35(2): 233–63.

Pettis, Michael. 2017. "Is Peter Navarro Wrong on Trade?" *China Financial Markets*, February 2.

Piskorski, Tomasz, Amit Seru, and James Witkin. 2015. "Asset Quality Misrepresentation by Financial Intermediaries: Evidence from the RMBS Market." *Journal of Finance* 70(6): 2635–78.

Ríos-Rull, José-Víctor, and Zhen Huo. 2016. "Financial Frictions, Asset Prices, and the Great Recession." Staff Report 526, Federal Reserve Bank of Minneapolis.

Ronald I. McKinnon. 1984. "The International Capital Market and Economic Liberalization in LDCs." *Developing Economics* 22: 478–81.

Schmitt-Grohé, Stephanie, and Martín Uribe. 2016. "Downward Nominal Wage Rigidity, Currency Pegs, and Involuntary Unemployment." *Journal of Political Economy* 124(5): 1466–1514.

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

Sufi, Amir. 2015. "Out of Many, One? Household Debt, Redistribution and Monetary Policy during the Economic Slump." Andrew Crockett Memorial Lecture, BIS.

Summers, Lawrence H. 2014. "US Economic Prospects: Secular Stagnation, Hysteresis, and the Zero Lower Bound." *Business Economics* 49(2): 65–73.

Verner, Emil, and Gyözö Gyöngyösi. 2017. "Household Debt Revaluation and the Real Economy: Evidence from a Foreign Currency Debt Crisis." Unpublished paper.

Wolf, Martin. 2014. *The Shifts and the Shocks: What We've Learned–and Have Still to Learn–from the Financial Crisis.* Penguin Books.

Yagan, Danny. 2017. "Employment Hysteresis from the Great Recession." NBER Working Paper 23844.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 59–86

Identification in Macroeconomics

Emi Nakamura and Jón Steinsson

Any scientific enterprise needs to be grounded in solid empirical knowledge about the phenomenon in question. Many of the main empirical questions in macroeconomics are the same as they have been since at least the Great Depression. What are the sources of business cycle fluctuations? How does monetary policy affect the economy? How does fiscal policy affect the economy? Why do some countries grow faster than others? Those new to our field may be tempted to ask, "How can it be that after all this time we don't know the answers to these questions?"

The reason is that identification in macroeconomics is difficult. Take the case of monetary policy. Unfortunately for us as empirical scientists, the Federal Reserve does not randomize when setting interest rates. Quite to the contrary, the Federal Reserve employs hundreds of PhD economists to pore over every bit of data about the economy so as to make monetary policy as endogenous as it possibly can be. This fact means that quite a bit of ingenuity and careful research is required to identify a component of monetary policy that is plausibly exogenous to future output and can, thus, be used to measure directly the effects of policy on output.

An important strand of empirical work in macroeconomics attempts the challenging task of identifying plausibly exogenous variation in macroeconomic policy and using this variation to assess the effects of the policy. We refer to this type of work as direct causal inference. Later in this article, we provide a critical assessment of several of the main methods that have been used to assess the effects of macroeconomic policies

■ *Emi Nakamura and Jón Steinsson are both Chancellor's Professors of Economics at the University of California at Berkeley, Berkeley, California. Their email addresses are enakamura@berkeley.edu and jsteinsson@berkeley.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.59 doi=10.1257/jep.32.3.59

in the academic literature—including identified vector autoregression. We do this in the context of asking what the best available evidence is on monetary nonneutrality.

A serious challenge faced by researchers attempting direct causal inference in macroeconomics is that the natural experiments we can find in the data are rarely exactly the experiments we would need to answer the policy questions in which we are interested. This "external validity" problem can be illustrated by thinking about monetary and fiscal policy. One issue is that the dynamic nature of monetary and fiscal policy makes these policies very high dimensional. Some monetary policy announcements only affect expectations about policy in the very short run (for example, whether the Federal Reserve will tighten this month or next month), while others affect policy expectations both in the short run and longer run, and still others only affect expectations about policy several years in the future (for example, when the policy interest rate is at the zero lower bound and the Fed makes a commitment to keep it there for longer than previously expected). The same is true of fiscal policy. The recent theoretical literature has emphasized that the future time profile of a policy action greatly affects its impact on current output and inflation. Identifying the effects of a policy shock with one time profile, therefore, does not necessarily identify the effects of a policy shock with a different time profile. A second external validity issue is that the effects of fiscal shocks depend on the response of monetary policy (for example, whether it is constrained by the zero lower bound). The effects of monetary policy, of course, also depend on the response of fiscal policies. A third issue is that the effects of monetary and fiscal policy may differ depending on the level of slack in the economy and how open the economy is. A fourth issue is that the degree to which a policy action is a surprise can affect both how strongly and when the economy reacts to it.

These external validity issues (and others) mean that even very cleanly identified monetary and fiscal natural experiments give us, at best, only a partial assessment of how *future* monetary and fiscal policy actions—which may differ in important ways from those in the past—will affect the economy. One response to these issues is to gather direct causal evidence about each and every different case. This however, may not be feasible. Even if it is feasible, it seems that one should be able to learn something about one case from evidence on another.

Due to the challenges described above, much empirical work in macroeconomics is more structural in nature. Such work often takes the form of researchers focusing on a set of moments in the data and arguing that these moments can discriminate between different models of how the economy works. Estimates of causal effects (that is, the response to identified structural shocks) can play an important role in this type of inference. The causal effects estimates can be viewed as target moments that models should match in much in the same way as unconditional means, variances, and covariance. We will use the term "identified moments" as a shorthand for "estimates of responses to identified structural shocks"—what applied microeconomists would call "causal effects."[1]

[1] The term "identified moments" may seems odd to some. In econometrics, it is parameters that are identified, not moments. We use the term "moment" in a broad sense to refer to a target statistic that

We argue that identified moments are often particularly informative moments for distinguishing between important classes of macroeconomic models. As a first example, consider the case of distinguishing between real business cycle and New Keynesian models—which is important for many policy questions. One approach is to use full information structural estimation methods such as maximum likelihood or Bayesian methods. Another approach is to match unconditional means, variances, and covariances in the tradition of the real business cycle calibration literature. In contrast, Rotemberg and Woodford (1997) and Christiano, Eichenbaum, and Evans (2005) estimate the response of output and inflation to identified monetary policy shocks and use these responses to discriminate between different business cycle models. Similarly, Galí (1999) and Basu, Fernald, and Kimball (2006) use the response of output and hours to identified productivity shocks to distinguish between models.

A second example is the recent literature on regional fiscal multipliers, and more generally the growing literature that aims to shed light on macroeconomic questions using cross-sectional identification strategies. Estimates of the regional government spending multiplier do not *directly* answer the policy question macroeconomists are most interested in—the effect of fiscal stimulus at the national level. However, the regional fiscal multiplier turns out to have a great deal of power in distinguishing between different models of the business cycle (Nakamura and Steinsson 2014). Models that can match a large regional multiplier typically imply that output responds strongly to demand shocks. In these models the aggregate multiplier is large when monetary policy is accommodative (for example, at the zero lower bound). For this reason, the recent literature on the regional fiscal multiplier has been able to provide powerful indirect evidence on the effectiveness of aggregate fiscal stimulus.

A third example is the use of estimates of the marginal propensity to consume (MPC) from a transitory fiscal rebate to distinguish between competing models of consumption dynamics. One approach uses truly random variation in the timing of fiscal stimulus checks to estimate a quarterly MPC of roughly 0.25 for nondurable consumption (for example, Johnson, Parker, Souleles 2006; Parker, Souleles, Johnson, and McClelland 2013). Kaplan and Violante (2014) use these estimates to distinguish between competing models of consumption dynamics. Their favored model adds illiquid assets that earn high returns to an otherwise standard model with uninsurable income risk and borrowing constraints. Angeletos, Laibson, Rebetto, Tobacman, and Weinberg (2001) argue that models in which households face self-control problems can help match the estimated MPC in the data.

The identifying assumptions that identified moments rely on are typically controversial. So, why use such moments? What is the upside of this approach relative to, for example, targeting simple unconditional moments? An important advantage

a researcher wants the model to match. We use the term "identified" because the target statistics we have in mind are estimates of causal effect parameters—or what macroeconomists would call estimated responses to "identified structural shocks"—as opposed to simple unconditional moments such as means, variances, and covariances.

is that in some cases the ability of a model to match identified moments may depend primarily on a particular sub-block of the model and be relatively insensitive to model misspecification in other sub-blocks of the model.[2] In the examples above, one does not, for example, have to take a stand on whether business cycles are driven by demand or supply shocks. In contrast, inference based on unconditional moments is typically highly sensitive to this. Moreover, Kaplan and Violante's (2014) inference based on the marginal propensity to consume seems unlikely to depend heavily on which frictions are included in the model outside of the consumption block, for example, price or wage rigidity, or investment and labor adjustment costs.

The Power of Portable Statistics

One important innovation of the early real business cycle literature was a move away from using likelihood-based empirical methods towards empirical evaluation based on matching moments (Kydland and Prescott 1982; Prescott 1986). An advantage of this approach is that it leads to the creation of "portable statistics" that can be used over and over again by researchers to discipline and test models. This allows for a division of labor and a fruitful back-and-forth between the theoretical and empirical parts of the field (and other fields). The equity premium is a good example of a portable statistic. Mehra and Prescott (1985) consider whether the equity premium is consistent with one specific class of models. A generation of subsequent researchers has then used this same statistic to evaluate a host of new models. The result has been an enormously influential literature on the nature of risk and risk aversion.[3]

It is useful to distinguish several types of moments that have been influential in empirical macroeconomics. We first consider the distinction between "micro moments" and "macro moments," and then the distinction between what we call "identified moments" and simpler moments.

Micro and Macro Moments

Micro moments are constructed using microeconomic data on the behavior of individuals and firms. A prominent example is the frequency of price change and related statistics on price rigidity (Bils and Klenow 2004; Nakamura and Steinsson 2008; Klenow and Kryvtsov 2008). These statistics help to discipline models that

[2] Our argument for moment matching using identified moments is related to Chetty's (2009) argument for sufficient statistics. However, we argue that identified moments can help answer many policy questions by distinguishing between models—that is, by learning about deep structural parameters. Chetty's emphasis is on combining identified moments into a formula (a "sufficient statistic") that answers a particular policy question.

[3] What is an example of a statistic that is not portable? The score of the likelihood function of a particular model (the moment that maximum likelihood estimation seeks to match) is very informative about that particular model. But it is not a very intuitive statistic to use to evaluate other models and is rarely (if ever) reused as a moment that other researchers (with other models) seek to match.

are designed to understand the effects of monetary policy. Another prominent example is the change in time spent shopping as well as the quantity and quality of food intake at the time of retirement (Aguiar and Hurst 2005). These statistics help distinguish between competing life-cycle models of household consumption and savings behavior.

Macro moments use aggregated data to identify equilibrium outcomes that are informative about what type of world we live in. The equity premium is an example of a highly influential macro moment, as are facts about changes in real wages and hours worked per person over the past century. The fact that real wages have risen by a large amount while hours worked have been stable or fallen slightly strongly rejects models without income effects on labor supply and, in fact, suggests that income effects are slightly larger than substitution effects in the long run. This motivates the use of "balanced growth preferences" in macroeconomic models (King, Plosser, and Rebelo 1988; Boppart and Krusell 2016).

There is a rich tradition in macroeconomics of using simple micro and macro moments to make inferences about how the world works. In many cases, these types of statistics can yield powerful inference. Prominent examples in addition to those discussed above include the real business cycle literature (Kydland and Prescott 1982; King and Rebelo 1999), the Shimer (2005) puzzle literature, the misallocation literature (Hsieh and Klenow 2009), the literature on exchange rate disconnect (Meese and Rogoff 1983; Itskhoki and Mukhin 2017), and the literature on "wedges" (Chari, Kehoe, and McGrattan 2008; Shimer 2009).

Causal Effects as Identified Moments

Here, we contrast simple statistics such as means, variances, and covariances with more complex statistics derived from empirical strategies designed to uncover what applied microeconomists would call causal effects, but macroeconomists would call responses to structural shocks. The last quarter century has seen a "revolution of identification" in many applied fields of economics (Angrist and Pischke 2010). This revolution has increased emphasis on identifying causal effects using credible research designs based on the use of instrumental variables, difference-in-difference analysis, regression discontinuities, and randomized controlled trials. It is these types of causal effects estimates that we refer to as identified moments.

In some cases, there is a one-to-one mapping between identified moments and a deep structural parameter. For example, there is a large literature in labor economics that estimates the labor supply elasticity (Chetty 2012; Chetty, Guren, Manoli, and Weber 2013). Macroeconomists have long made use of causal effects estimates of this kind to discipline the models that they work with. In the jargon of macroeconomics, we frequently "calibrate" certain parameters of our models (such as the labor supply elasticity) based on external estimates.

Many identified moments, however, do not correspond directly to a deep structural parameter. Two prominent examples we discussed earlier are estimates of the marginal propensity to consume out of a transitory fiscal rebate and estimates of the regional fiscal multiplier. In these cases, a theoretical framework is required to go from the identified moment to the macroeconomic questions of interest. These types of identified moments are valuable because they can be used as empirical targets in a structural moment-matching exercise aimed at distinguishing between competing models that differ in their implications about the macroeconomic question of interest.[4]

There is a prevalent view in macroeconomics that if your empirical strategy is to calculate the same moment in real-world data as in data from a set of models, you might as well focus on very simple unconditional moments. A moment is a moment, the argument goes. But relative to simple unconditional moments, the advantage of identified moments is that they can provide evidence on specific causal mechanisms of a model and may be relatively invariant to other model features.

Consider the recent debate on the role of changes in house prices in causing the Great Recession of 2007–2009. Mian, Rao, and Sufi (2013) and Mian and Sufi (2014) compare changes in consumption and employment in metropolitan areas that experienced larger or smaller house price changes. Of course, causation may run both ways: increases in house prices may stimulate economic activity, but a local boom may also increase house prices. To isolate the causal effect of house prices on consumption and employment, these authors propose to instrument for changes in house prices with estimates of housing supply elasticities constructed by Saiz (2010), which in turn are constructed from data on regional topology and land-use regulation. The idea for identification is that national shocks will lead house prices to increase more in metropolitan areas where housing supply is less elastic.[5] Using this identification strategy, Mian, Rao, and Sufi (2013) find that the elasticity of consumption with respect to housing net worth is 0.6 to 0.8, while Mian and Sufi (2014) find that the elasticity of nontradable employment with housing net worth is 0.37.

These estimates do not directly answer the macroeconomic question of how much aggregate house prices affect economic activity because this empirical approach is based on comparing one metropolitan area to another and therefore "differences out" aggregate general equilibrium effects. However, these estimates are quite informative about the "consumption block" of macroeconomic models. They strongly reject simple complete-markets models of consumption such as the influential model of Sinai and Souleles (2005)—in which the elasticity of consumption to house prices is zero—in favor of models with life-cycle effects, uninsurable income risk, borrowing constraints, and in which households can substitute away from housing when its price rises (for a discussion of such a model, see Berger, Guerrieri, Lorenzoni, and Vara 2018).

[4] Formally, the idea is to use indirect inference with limited-information empirical models designed to estimate causal effects (for example, an instrumental variables regression, a difference-in-difference design, or a regression-discontinuity design) as auxiliary models. See Smith (2008) for an introduction to indirect inference.

[5] The housing supply elasticity is obviously not randomly assigned (for example, land availability is correlated with whether a city is on the coast). However, the argument for identification is that whatever makes these coastal (and otherwise land-constrained) locations different does not affect their response to aggregate shocks directly, except through the implications for the housing market.

The identification strategy used by Mian, Rao, and Sufi (2013) and Mian and Sufi (2014) is by no means uncontroversial, as is often the case with identified moments. However, focusing instead on simple moments like the raw correlation between house prices and consumption or house prices and employment has downsides too. These simple moments are likely to be sensitive to assumptions that have little to do with consumption behavior, such as what shocks drive the business cycle, and to the strength of general equilibrium effects, such as the response of prices and wages to demand shocks (which, in turn, may depend on virtually everything about how the model works). In other words, using these simple moments results in a joint test of all the parameters and assumptions in the model, while identified moments can focus in on the consumption block of the model and provide inference that is robust to the specification of other parts of the model.[6]

Another prominent example of an identified moment matching exercise is recent work that seeks to determine the role of unemployment insurance extensions in delaying recovery from the Great Recession. Hagedorn, Manovskii, and Mitman (2015) and Chodorow-Reich, Coglianese, and Karabarbounis (2017) estimate the response of unemployment to variation in unemployment benefit extensions across states using discontinuity-based identification and an instrumental variables approach, respectively. They then use these cross-sectional identified moments to determine key parameters in a labor market search model and use the resulting model to determine how unemployment insurance extensions will affect the economy at the aggregate level. Here, again, it is likely possible to pin down these same parameters using simple moments (such as the variance and covariance of unemployment, vacancies, wages, and benefit extensions) in a fully specified structural model. However, such an exercise would likely be sensitive to many auxiliary features of the model being used outside the "labor market block" of the model.

In the examples above, identified moments provide information primarily on a particular "block" or mechanism of a macroeconomic model. This "piecemeal" form of inference will, therefore, result in partial identification on the model space. It is inevitable that any given statistic or set of statistics will not be able to pick out a single model and reject all others. The fact that several models are consistent with a statistic is not grounds for rejecting the statistic as being uninteresting. We should instead think in reverse: If a statistic has power to reject an important set of models in favor of another set of models, the statistic is useful.

[6] Similar methods have been applied to assess the importance of financial frictions on firms. Catherine, Chaney, Huang, Sraer, and Thesmar (2017) use instrumental variable estimates of the response of firm investment to changes in real estate collateral to pin down parameters in a structural model with financial frictions and then use the model to quantify the aggregate effects of relaxing collateral constraints. Chodorow-Reich (2014) uses bank shocks to identify the effects of financial constraints on firm employment. In an appendix, he uses these micro-level estimates to pin down parameters in a general equilibrium model of the effect of bank shocks on the economy as a whole. Huber (2018) estimates direct and indirect firm effects as well as regional effects of a large bank shock in Germany. Other important papers in this literature include Peek and Rosengren (2000), Calomiris and Mason (2003), Ashcraft (2005), Greenstone, Mas, and Nguyen (2017), and Mondragon (2018).

Aggregate versus Cross-Sectional Identification

The increased use of cross-sectional identification approaches has been an exciting development in empirical macroeconomics. In this work, researchers use geographically disaggregated panel datasets—often disaggregated to the level of the state or metropolitan statistical area—to identify novel causal effects. The use of regional data typically multiplies the number of data points available by an order of magnitude or more. It also allows for difference-in-difference identification and makes possible the use of a powerful class of instrumental variables: differential regional exposure to aggregate shocks. Prominent examples of this approach include: Mian and Sufi (2014) on the role of the housing net worth channel in the Great Recession; Autor, Dorn, and Hanson (2013) on the effects of Chinese imports on US employment; Beraja, Hurst, and Ospina (2016) on wage rigidity during the Great Recession; Martin and Philippon (2017) on the effects of debt during the Great Recession in the eurozone; Carvalho et al. (2016) on supply chain disruptions after the Great East Japan Earthquake of 2011; and a large literature on fiscal stimulus discussed below.

A key challenge for this literature is how to translate regional responses into aggregate responses. A common approach is to add up regional responses with the implicit assumption that the least affected region is unaffected by the shock and report this sum as the aggregate response. However, this approach ignores general equilibrium effects that influence the aggregate response but are absorbed by time fixed effects in the cross-sectional regressions used in this literature. As an example, Mian and Sufi show that the dramatic fall in house prices between 2006 and 2009 did not *differentially* affect tradables employment in areas with larger house price declines. However, this does not mean that this shock had no effect on tradables employment in the aggregate. (To be clear, they don't make any such claim.) Typically, regional responses can only be translated into aggregate responses through the lens of a fully specified general equilibrium model (Nakamura and Steinsson 2014).

A common critique of estimates based on cross-sectional identification in macroeconomics is that they don't answer the right question. While it is true that these estimates don't *directly* provide estimates of aggregate responses, they often provide a great deal of indirect evidence by helping researchers discriminate between different theoretical views of how the world works. In the language of the previous section, these cross-sectional estimates are examples of identified macro moments. They can be used as moments in a moment-matching exercise that is aimed at distinguishing between important classes of general equilibrium structural models of the economy that have different implications about the primary question of interest. This combination of theory and empirics can yield very powerful inference. The literature on the stimulative effects of government spending provides a nice case study to illustrate these ideas.

Fiscal Stimulus: Aggregate Evidence

Direct aggregate evidence on the government spending multiplier is far from conclusive. This evidence largely comes in two forms: evidence from wars and from vector autoregressions. Barro and Redlick (2011) regress changes in output on changes in defense spending and a few controls for a US sample period including several major wars. Their main conclusion is that the government purchases multiplier is between 0.6 and 0.7. Virtually all of the identification in their sample comes from World Wars I and II, and to a lesser extent the Korean War. When they restrict attention to data from after the Korean War, the confidence interval for their estimate includes all remotely plausible values.

Barro and Redlick (2011) assume that war-related defense spending is exogenous to output. Conceptually, this is like using wars as an instrument for government spending. The strength of this approach is that reverse causation is not likely to be a problem. World War I, World War II, and the Korean War did not happen because the US economy was in a recession or a boom. However, for war-related spending to be exogenous, wars must only affect output through spending. This is unlikely to be true. Barro and Redlick are aware of this issue and discuss some potential confounding factors. On one side, patriotism likely increased labor supply during major wars and thus results in an upward bias of the multiplier. On the other hand, wartime rationing and price controls likely result in a downward bias. Barro and Redlick argue that patriotism is likely the dominant bias, while Hall (2009) argues that the effects of wartime controls result in a net downward bias.

Blanchard and Perotti (2002) regress government spending on four quarterly lags of itself, taxes, output, a quadratic time trend, and a dummy variable for 1975:Q2. They view the residual from this regression as exogenous shocks to government spending. [7] They construct an impulse response function for output, consumption, and other variables to the government spending shocks by iterating forward a vector autoregression. Their baseline sample period is 1960–1997. The estimated response of output swings up, then down, then up again, with a peak response of output of 1.3 times the initial response of spending 15 quarters after the impulse.

Blanchard and Perotti's (2002) approach to identification makes the strong assumption that controlling for four lags of taxes, spending, and output eliminates all endogenous variation in spending. They argue that this approach to identification is more plausible for fiscal policy than for monetary policy, because output stabilization is not as dominant a concern of fiscal policy and because implementation lags are longer in fiscal policy. Ramey (2011) argues, however, that Blanchard and Perotti miss an important part of the response to fiscal shocks because news about fiscal shocks—especially those associated with major wars—arrives well ahead of the main increase in spending. Another concern is that estimates of the fiscal multiplier based on post–World War II aggregate data—such as Blanchard and Perotti's estimates—have such large standard errors that few interesting hypotheses can be rejected. Also, their estimates are highly sensitive to the sample period (like whether the Korean War is included) and to which controls are included (Galí, López-Salido, and Vallés 2007; Ramey 2016).

[7] This methodology is equivalent to performing a Cholesky decomposition of the reduced form errors from the vector autoregression with government spending ordered first.

Finally, aggregate estimates of the government spending multiplier are subject to an important external validity problem having to do with the response of monetary policy. Blanchard and Perotti's estimates come from a time period when monetary policy could "lean against the wind" by raising real interest rates in response to a government spending shock. The aggregate fiscal multiplier is potentially quite different in the midst of a deep recession when monetary is constrained by the zero lower bound on nominal interest rates. Because fiscal stimulus packages tend to be discussed in recessions, evidence on the government spending multiplier for this circumstance is particularly valuable. Unfortunately, direct aggregate evidence on the effectiveness of government spending when monetary policy is constrained is even less well established than on the simpler question of the average aggregate multiplier (Ramey and Zubairy 2018; Miyamoto, Nguyen, and Sergeyev 2018).

An indirect way to infer the effectiveness of government spending when monetary policy is constrained is to amass evidence about whether a New Keynesian or neoclassical model is a better description of the world. In the New Keynesian model, the aggregate multiplier from government spending can be quite low if monetary policy is responsive, but when monetary policy is unresponsive—say, at the zero-lower-bound—the aggregate multiplier can be quite large. In contrast, in a neoclassical model, the aggregate multiplier is small independent of monetary policy. Unfortunately, evidence on the aggregate fiscal multiplier is not very helpful in this regard. Estimates between 0.5 and 1.0—which is where most of the more credible estimates based on US data lie—are consistent with both of these models. As we explain below, cross-sectional evidence on the effects of government spending yields much sharper inference on this point.

Fiscal Stimulus: Cross-Sectional Evidence

In recent years, researchers have used a wide array of cross-sectional identification strategies to estimate the effects of government spending. Examples include windfall returns on state pension plans (Shoag 2015), differential state sensitivity to military buildups (Nakamura and Steinsson 2014), crackdowns on Mafia-infiltrated municipalities in Italy (Acconcia, Corsetti, and Simonelli 2014), formulas used to allocate spending across states from the American Recovery and Reinvestment Act (Chodorow-Reich, Feiveson, Liscow, and Woolston 2012; Wilson 2012; Dupor and Mehkari 2016), and spending discontinuities associated with decadal population estimate revisions (Suárez Serrato and Wingender 2016). Chodorow-Reich (2017) surveys this literature in detail.[8]

Estimates of the regional spending multiplier from this literature tend to cluster in the range of 1.5–2.0, which is substantially larger than typical estimates of the aggregate multiplier. However, these two sets of estimates are not necessarily inconsistent. After all, they measure different things. To understand this point, consider the identification strategy in Nakamura and Steinsson (2014). The basic

[8] Nekarda and Ramey (2011) consider variation in government spending across industries, as opposed to variation across regions as in the literature discussed above.

idea is that national military buildups (like the Carter–Reagan buildup following the Soviet invasion of Afghanistan in 1979) result in much larger changes in military spending in some states than others because the plants that build military hardware are unevenly distributed across the country. For example, national military buildups imply that spending rises much more in California than in Illinois. We then ask whether this translates into bigger increases in output in California than in Illinois. Our conclusion is that output in California rises by roughly $1.5 relative to output in Illinois for each extra $1 of spending in California relative to Illinois.

Importantly, our specification includes time fixed effects, which implies that our multiplier estimates are only identified off of the response of California relative to Illinois, not the response of all states to the aggregate buildup. This also means that aggregate general equilibrium effects are absorbed by the time fixed effects. These include any tightening of monetary policy that may occur as a consequence of the military buildup and the change in federal taxes needed to finance the buildup. Estimates of aggregate multipliers include these effects. Regional multiplier effects therefore do not provide direct evidence on the aggregate multiplier.

However, the regional multiplier provides a powerful diagnostic tool for distinguishing between competing macroeconomic models, and thereby for indirectly learning about the effectiveness of aggregate fiscal stimulus. In Nakamura and Steinsson (2014), we write down several multi-region business cycle models, simulate the same policy experiment in these models as we identify in the data, and compare the regional multipliers generated by each model with estimates of regional multipliers from real-world data. The textbook real business cycle model generates regional multipliers that are substantially smaller than our empirical estimate. In this model, a "foreign" demand shock (the federal government demanding military goods), leads people to cut back on work effort in other areas, which implies that the regional multiplier is less than one. We conclude that the regional multiplier evidence favors models in which output responds more strongly to a foreign demand shock than this model implies. We present an example of a Keynesian model in which the aggregate multiplier can be large (for example, when monetary policy is constrained at the zero lower bound). The regional multiplier does not uniquely identify a correct model (and no single statistic will). However, large regional multiplier estimates suggest that researchers should put more weight on models in which demand shocks can have large effects on output.

Monetary Policy: What Is the Best Evidence We Have?

What is the most convincing evidence for monetary nonneutrality? When we ask prominent macroeconomists this question, the three most common answers have been: the evidence presented in Friedman and Schwartz (1963) regarding the role of monetary policy in the severity of the Great Depression; the Volcker disinflation of the early 1980s and accompanying twin recession; and the sharp break in the volatility of the US real exchange rate accompanying the breakdown of the Bretton

Woods system of fixed exchange rates in 1973 (first emphasized by Mussa 1986).[9] It is interesting that two of these pieces of evidence—the Great Depression and Volcker disinflation—are large historical events often cited without reference to modern econometric analysis, while the third is essentially an example of discontinuity-based identification. Conspicuous by its absence is any mention of evidence from vector autoregressions, even though such methods have dominated the empirical literature for quite some time. Clearly, there is a disconnect between what monetary economists find convincing and what many of them do in their own research.

Large Shocks

The holy grail of empirical science is the controlled experiment. When it comes to monetary policy, for obvious reasons, we cannot do controlled experiments. We must instead search for "natural experiments"—that is, situations in which we can argue that the change in policy is large relative to potential confounding factors. Much empirical work takes the approach of seeking to control for confounding factors as well as possible. A different approach is to focus on large policy actions for which we can plausibly argue that confounding factors are drowned out. Such policy actions are, of course, rare. But looking over a long period and many countries, we may be able to piece together a body of persuasive evidence on the effects of monetary policy.

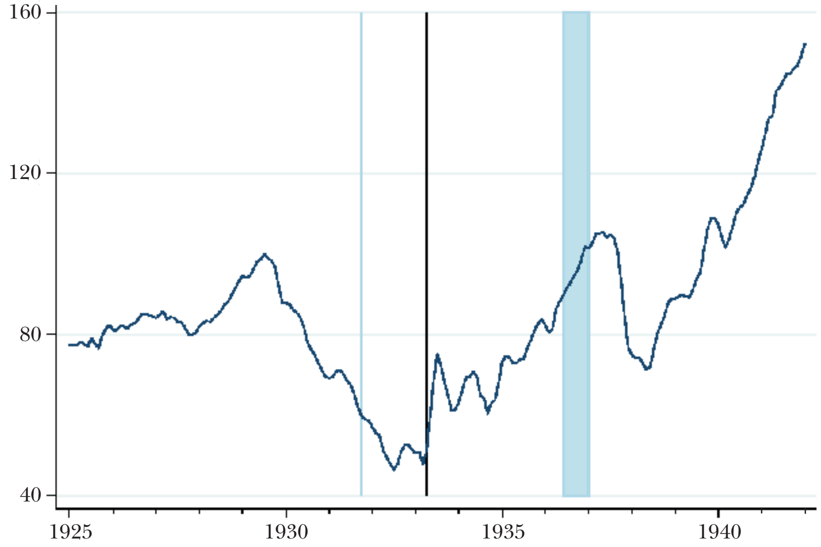
Friedman and Schwartz (1963, p. 688) argue in the final chapter of their monumental work on US monetary history that three policy actions taken by the Federal Reserve in the interwar period were 1) "of major magnitude" and 2) "cannot be regarded as necessary or inevitable economic consequences of contemporary changes in money income and prices." They furthermore argue that "like the crucial experiments of the physical scientist, the results are so consistent and sharp as to leave little doubt about their interpretation." The dates of these events are January–June 1920, October 1931, and July 1936–January 1937. We will focus on the latter two of these events, which occurred during the Great Depression.

Figure 1 plots the evolution of US industrial production from 1925 to 1942. The fall from the July 1929 peak to March 1933 was a staggering 53 percent. The recovery between March 1933 and the subsequent peak in May 1937 was rapid and large. But then a second very large downturn occurred. From May 1937 to May 1938, industrial production fell by 33 percent.

The light vertical bars in Figure 1 show the times of the two policy mistakes that Friedman and Schwartz (1963) highlight. In October 1931, the Federal Reserve raised the rediscount rate (the policy rate of that time) sharply from 1.5 to 3.5 percent in response to a speculative attack on the US dollar that followed Britain's decision to leave the gold standard. The Fed drastically tightened policy despite the fact that industrial production was in free fall and a wave of bank failures was underway. At first pass, it may seem reasonable to interpret this as a clean

[9] Of course, a significant fraction say something along the lines: "I know in my bones that monetary policy has no effect on output."

Figure 1
Industrial Production from 1925 to 1942
(index equals 100 in July 1929)



Note: The figure plots an index for industrial production in the US economy from January 1925 to January 1942. The index is equal to 100 in July 1929 (the peak month). The shaded bar and vertical line are the periods of policy mistakes identified by Friedman and Schwartz (1963). The dark vertical line is the time at which Roosevelt took the United States off the gold standard.

monetary shock. However, the subsequent fall in industrial production is not very different from the fall in the previous two years. It is not clear how much of the subsequent fall in industrial production is due to this policy shock as opposed to other developments that led to the equally rapid fall in the previous two years.

The second monetary shock emphasized by Friedman and Schwartz is more promising in this regard. From July 1936 to January 1937, the Fed announced a doubling of reserve requirements (fully implemented by May 1937) and the Treasury engaged in sterilization of gold inflows. Before this period, industrial production had been rising rapidly. Shortly after, it plunged dramatically. Friedman and Schwartz argue that the Fed's policy actions caused this sharp recession. However, a closer look reveals important confounding factors. Fiscal policy tightened sharply in 1937 because of the end of the 1936 veterans' bonus and the first widespread collection of Social Security payroll taxes, among other factors. In fact, prior to Friedman and Schwartz's (1963) work, Keynesians often held up the 1937 recession as an example of the power of fiscal policy. Romer and Romer (1989) also emphasize that 1937 was a year of substantial labor unrest. For these reasons, it is perhaps not clear that this episode is "so consistent and sharp as to leave little doubt about [its] interpretation."

A more general argument runs through the Friedman and Schwartz (1963) narrative of the Great Depression period. They argue that the Fed failed to act from early 1930 and March 1933, and instead allowed the money supply to fall and a substantial fraction of the banking system to fail. Eichengreen (1992) argues that an important reason why the Fed did not act during this period was that effective action would have been inconsistent with remaining on the gold standard. One of President Franklin Roosevelt's first policy actions was to take the United States off the gold standard in April 1933, shown by the black vertical line in Figure 1. The dollar quickly depreciated by 30 percent. Industrial production immediately skyrocketed. But, of course, the Roosevelt administration changed a number of policies. Whether it was going off gold or something else that made the difference is not entirely clear.[10]

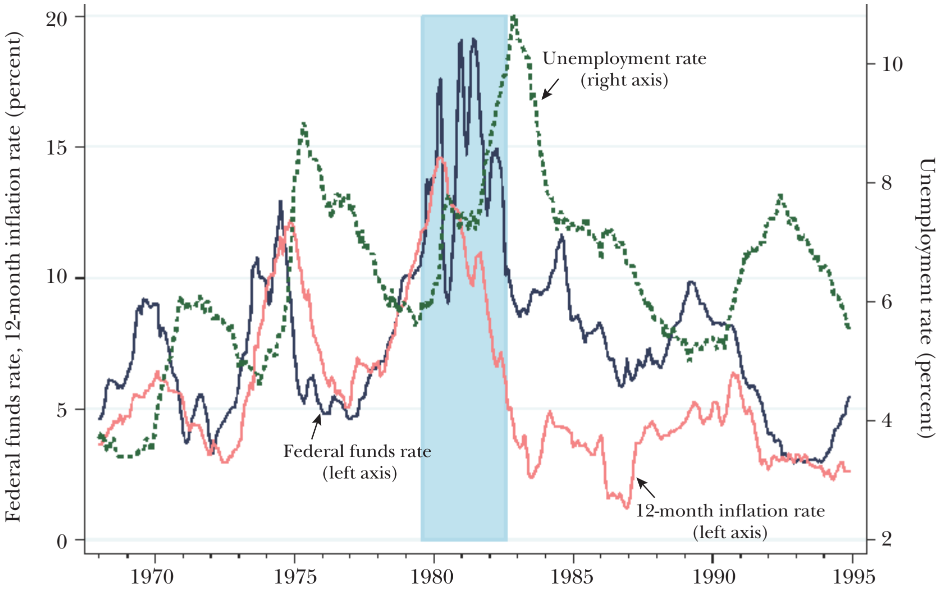
The Volcker disinflation and accompanying twin recessions of the late 1970s and early 1980s are another often-cited piece of evidence on monetary nonneutrality. US inflation had been low and stable since the end of the Korean War in the early 1950s, but then started to rise in the late 1960s. In the 1970s, inflation was both higher and much more volatile than before, as shown in Figure 2. Monetary policy during the 1970s is often described as "stop-go"—tight when the public was exercised about inflation and loose when the public was exercised about unemployment (Goodfriend 2007).

In August 1979, Paul Volcker became chairman of the Federal Reserve. Under Volcker's leadership, policy interest rates rose dramatically between October 1979 and March 1980. However, as shown in Figure 2, the Fed then eased policy such that rates fell even more dramatically (by 9 percentage points) in the spring and summer of 1980 as it became clear that the economy was contracting strongly. In the fall of 1980, inflation was no lower than a year earlier, and the Fed's credibility was, if anything, worse than before. Goodfriend and King (2005) argue that it was only at this point—in November 1980—that Volcker truly broke with prior behavior of the Fed and embarked on a sustained, deliberate disinflation. Interest rates rose dramatically to close to 20 percent and the Fed kept policy tight even in the face of the largest recession the US had experienced since the Great Depression.

The behavior of output during this period is consistent with the view that monetary nonneutrality is large. Output fell dramatically in the spring and summer of 1980 shortly after the Fed raised interest rates sharply. Output then rebounded in late 1980 shortly after the Fed reduced interest rates sharply. Output then fell by a large amount for a sustained period in 1981–1982 while the Fed maintained high interest rates to bring down inflation. Finally, output started recovering when the Fed eased monetary policy in late 1982.

[10] Eichengreen and Sachs (1985) offer related evidence that supports a crucial role for going off gold; they show that countries that went off the gold standard earlier, recovered earlier from the Great Depression. Also, Eggertsson and Pugsley (2006) and Eggertsson (2008) present a model and narrative evidence suggesting that the turning points in 1933, 1937, and 1938 can all be explained by a commitment to reflate the price level (1933 and 1938) and an abandonment of that commitment (1937). Going off gold was an important element of this commitment in 1933.

Figure 2
Federal Funds Rate, Inflation, and Unemployment from 1965 to 1995



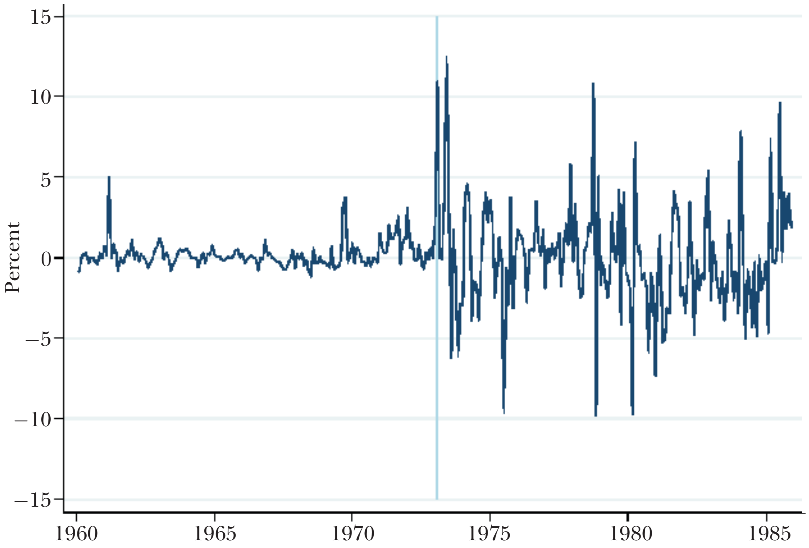
Note: The figure plots the federal funds rate (dark solid line, left axis), the 12-month inflation rate (light solid line, left axis), and the unemployment rate (dashed line, right axis). The Volcker disinflation period is the shaded bar (August 1979 to August 1982).

Many economists find the narrative account above and the accompanying evidence about output to be compelling evidence of large monetary nonneutrality.[11] However, there are other possible explanations for these movements in output. There were oil shocks both in September 1979 and in February 1981 (described in Table D.1 in the online Appendix). Credit controls were instituted between March and July of 1980. Anticipation effects associated with the phased-in tax cuts of the Reagan administration may also have played a role in the 1981–1982 recession (Mertens and Ravn 2012).

While the Volcker episode is consistent with a large amount of monetary nonneutrality, it seems less consistent with the commonly held view that monetary policy affects output with "long and variable lags." To the contrary, what makes the Volcker episode potentially compelling is that output fell and rose largely in sync with the actions of the Fed. If not for this, it would have been much harder to attribute the movements in output to changes in policy.

[11] The Volcker disinflation has also had a profound effect on beliefs within academia and in policy circles about the ability of central banks to control inflation. Today the proposition that inflation is "always and everywhere a monetary phenomenon" is firmly established; so firmly established that it is surprising to modern ears that this proposition was doubted by many in the 1970s (Romer and Romer 2002; Nelson 2005).

Figure 3
Monthly Change in the US–German Real Exchange Rate



Note: The figure plots the monthly change in the US–German real exchange rate from 1960 to 1990. The vertical line marks February 1973, when the Bretton Woods system of fixed exchange rates collapsed.

Discontinuity-Based Identification

There is incontrovertible, reduced-form evidence that monetary policy affects relative prices. The evidence on this point is strong because it can be assessed using discontinuity-based identification methods. The pioneering paper on this topic is Mussa (1986). He argued that the abrupt change in monetary policy associated with the breakdown of the Bretton Woods system of fixed exchange rates in February 1973 caused a large increase in the volatility of the US real exchange rate. Figure 3 plots the monthly change in the US–German real exchange rate from 1960 to 1990. There is a clear break in the series in February 1973: its standard deviation rose by more than a factor of four. The switch from a fixed to a flexible exchange rate is a purely monetary action. In a world where monetary policy has no real effects, such a policy change would not affect real variables like the real exchange rate. Figure 3 demonstrates dramatically that the world we live in is not such a world.

As with any discontinuity-based identification scheme, the identifying assumption is that other factors affecting the real exchange rate do not change *discontinuously* in February 1973. The breakdown of the Bretton Woods system was caused by a gradual build-up of imbalances over several years caused by persistently more inflationary macroeconomic policy in the United States than in Germany, Japan, and other countries. There were intense negotiations for over a year before the system finally collapsed. It is hard to think of plausible alternative explanations for the *discontinuous* increase in the volatility of the US real exchange rate that occurred at this time.[12]

There is also strong discontinuity-based evidence that monetary policy affects real interest rates. A large amount of monetary news is revealed discretely at the time of the eight regularly scheduled meetings of the Federal Open Market Committee (FOMC) of the Federal Reserve. In Nakamura and Steinsson (forthcoming), we construct monetary shocks using changes in interest rates over a 30-minute window surrounding FOMC announcements. Over such a short window, movements in interest rates are dominated by the monetary announcement. Furthermore, if financial markets are efficient, any systematic response of the Fed to information about the economy that is public at the time of the announcement is already incorporated into financial markets, and, therefore, does not show up as spurious variation in the monetary shocks we construct. We show that nominal and real interest rates respond roughly one-for-one to these monetary shocks several years out into the term structure of interest rates, while expected inflation responds little. For example, the three-year nominal and real forward rates respond by very similar amounts. Hanson and Stein (2015) present similar empirical results.

Direct high-frequency evidence of the effect of monetary policy on *output* is much weaker than for relative prices (Cochrane and Piazzesi 2002; Angrist, Jordà, and Kuersteiner 2017). The reason is that high-frequency monetary shocks are quite small, as is often the case with very cleanly identified shocks. This implies that the statistical power to assess their effect on output several quarters in the future is limited (because many other shocks also affect output over longer time periods).

The effect of monetary shocks on output can be broken into two separate questions: First, how much do monetary shocks affect various relative prices? Second, how much do these various relative prices affect output? Viewing things this way helps clarify that it is the first question that is *the* distinguishing feature of models in which monetary policy has effects on output. All models—neoclassical and New Keynesian—imply that relative prices affect output. However, only in some models does monetary policy affect relative prices, and these models typically imply that money is nonneutral.

A complicating factor regarding the high-frequency evidence from announcements of the Federal Open Market Committee is that these announcements may lead the private sector to update its beliefs not only about the future path of monetary policy, but also about other economic fundamentals. If the Fed reveals greater optimism about the economy than anticipated, the private sector may revise its own beliefs about where the economy is headed. We find evidence for such "Fed information effects" in Nakamura and Steinsson (forthcoming). A surprise tightening of policy is associated with an *increase* in expected output growth in the Blue Chip survey of professional forecasters (the opposite of what standard analysis of monetary shocks

[12] Velde (2009) presents high-frequency evidence on the effects of a large monetary contraction in 18th-century France. He shows that domestic prices responded sluggishly and incompletely to this shock. Burstein, Eichenbaum, and Rebelo (2005) show that large, abrupt devaluations lead to large changes in real exchange rates that are mostly due to changes in the relative price of nontradable goods.

would imply). We show that a New Keynesian model that incorporates Fed information effects can match this fact as well as facts about the response of real interest rates and expected inflation. In this model, a Fed tightening has two effects on the economy: a traditional contractionary effect through increases in real interest rates relative to natural rates and a less traditional expansionary effect coming from the Fed's ability to increase optimism about the economy. Earlier evidence on Fed information effects is presented by Romer and Romer (2000), Faust, Swanson, and Wright (2004), and Campbell, Evans, Fisher, and Justiniano (2012).

The Fed information effect implies that an external validity problem arises whenever researchers use responses to monetary shocks to make inferences about the effects of systematic monetary policy actions. Surprise monetary actions lead to information effects, while the systematic response of the Fed to new data do not. This implies that surprise monetary shocks are less contractionary than the systematic component of monetary policy.

Using the Narrative Record to Identify Shocks

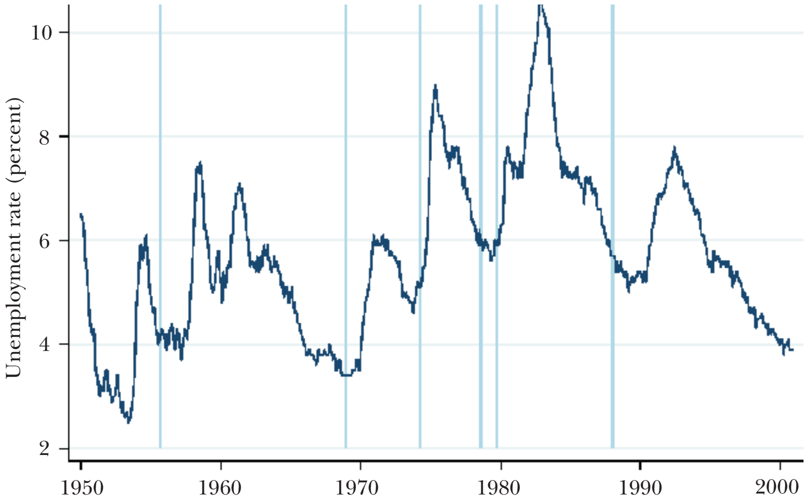
Romer and Romer (1989) argue that contemporaneous Federal Reserve records can be used to identify natural experiments. They use such records to identify "episodes in which the Federal Reserve attempted to exert a contractionary influence on the economy in order to reduce inflation" in the post-World War II period. They identify six such episodes and subsequently added a seventh (Romer and Romer 1994). Figure 4 shows that after each of the Romer–Romer dates, marked by vertical lines, unemployment rises sharply. Pooling the data from these seven episodes, Romer and Romer argue that together they constitute strong evidence for substantial real effects of monetary policy.

While this "narrative approach" to identification is clearly valuable, it faces several challenges. First, narrative shocks are selected by an inherently opaque process. This raises the concern that the results are difficult to replicate. Second, with only seven data points, it may happen by chance that some other factor is correlated with the monetary shocks. In cases when one has dozens or hundreds of shocks, any random correlation with some other factor is likely to average to zero. But with seven data points, this may not happen. In fact, Hoover and Perez (1994) argue that Romer and Romer's monetary dates are strikingly temporally correlated with dates of oil shocks (see Table D.1 in the appendix).

Third, narrative shocks are often found to be predictable, suggesting the possibility of endogeneity. In the case of Romer and Romer's (1989) analysis, this concern was raised by Shapiro (1994) and Leeper (1997). However, it can be difficult to convincingly establish predictability due to overfitting concerns. The cumulative number of regressions run by researchers trying to assess the predictability of narrative shocks may be very large. Even by chance, some of these should turn up estimates that are statistically significant.[13]

[13] In the case of Leeper's (1997) results, Romer and Romer (1997) present a simple pseudo-out-of-sample procedure demonstrating dramatic overfitting. They redo Leeper's analysis seven times, in each case leaving

Figure 4
Unemployment from 1950 to 2000



Note: The figure plots the unemployment rate from 1950 to 2000. The light vertical lines indicate the dates identified by Romer and Romer (1989, 1994) as "episodes in which the Federal Reserve attempted to exert a contractionary influence on the economy in order to reduce inflation."

Controlling for Confounding Factors

The most prevalent approach to identifying exogenous variation in monetary policy is to attempt to control for confounding factors. Much of the vector autoregression literature takes this approach. A common specification is to regress the federal funds rate on contemporaneous values of several variables (such as output and inflation), as well as several lags of itself and these other variables, and then to view the residuals from this regression as exogenous monetary policy shocks.[14]

This approach to identifying monetary policy shocks is often described as involving "minimal identifying assumptions." In our view, however, the implicit assumptions are very strong. What is being assumed is that controlling for a few lags of a few variables captures all endogenous variation in policy. This seems highly unlikely to be true in practice. The Fed bases its policy decisions on a huge amount of data. Different considerations (in some cases highly idiosyncratic) affect policy at different times. These include stress in the banking system, sharp changes in commodity prices, a recent stock market crash, a financial crisis in emerging markets, terrorist attacks, temporary investment tax credits, and the Y2K computer glitch. The list goes on and on. Each of these considerations may only affect policy in a meaningful way on a small number of dates, and the number of such influences is so large that it is not feasible

out one of the monetary shock dates. For each set of estimates, they then see if the model can predict the shock date that is left out. Using this pseudo-out-of-sample procedure, they find no predictability.

[14] A common way of describing this procedure is as performing a Cholesky decomposition of the reduced form errors from the vector autoregression with the federal funds rate ordered last.

to include them all in a regression. But leaving any one of them out will result in a monetary policy "shock" that the researcher views as exogenous but is in fact endogenous. Rudebusch (1998) is a classic discussion of these concerns.

To demonstrate this point, consider an example. Following Cochrane and Piazzesi (2002), Figure 5 plots the evolution of the federal funds rate target of the Federal Reserve as well as the one-month Eurodollar rate—that is, a one-month interbank interest rate on US dollars traded in London—around the time of the September 11, 2001, terrorist attacks. The Eurodollar rate may be viewed as the average expected federal funds rate over the next month. On September 10, the Eurodollar rate traded at 3.41 percent, quite close to the target federal funds rate of 3.5 percent. Markets did not open on September 11 due to the terrorist attacks in New York. Before markets reopened on September 17, the Fed announced a 50 basis points drop in the target federal funds rate to 3 percent and the one-month Eurodollar rate traded at 2.97 percent that day. The futures contract for the federal funds rate in September 2001 reveals that this 50 basis points drop in the federal funds rate was completely unanticipated by markets as of September 10. The one-month Eurodollar rate was trading below the target for the federal funds rate on September 10 because markets were anticipating an easing of 25–50 basis points at the regularly scheduled meeting of the Federal Open Market Committee on October 2.

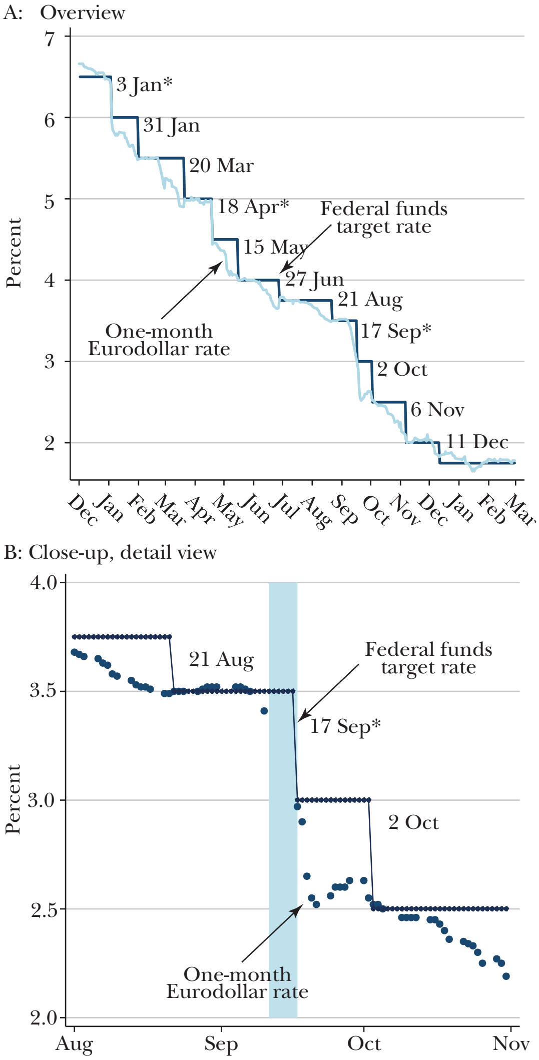
The easing on September 17 was obviously due to the terrorist attacks and therefore obviously endogenous: the terrorist attacks caused the Fed to revise its assessment about future growth and inflation, leading to an immediate drop in interest rates. However, standard monetary vector autoregressions treat this policy easing as an exogenous monetary shock. The reason is that the controls in the policy equation of the vector autoregression are not able to capture the changes in beliefs that occur on 9/11. Nothing has yet happened to any of the controls in the policy equation (even in the case of a monthly vector autoregression). From the perspective of this equation, therefore, the drop in interest rates in September 2001 looks like an exogenous easing of policy. Any unusual (from the perspective of the vector autoregression) weakness in output growth in the months following 9/11 will then, perversely, be attributed to the exogenous easing of policy at that time. Clearly, this is highly problematic.

In our view, the way in which identification assumptions are commonly discussed in the vector autoregression literature is misleading. It is common to see "the" identifying assumption in a monetary vector autoregression described as the assumption that the federal funds rate does not affect output and inflation contemporaneously. This assumption sounds innocuous, almost like magic: You make one innocuous assumption, and voilà, you can estimate the dynamic causal effects of monetary policy. We remember finding this deeply puzzling when we were starting off in the profession.

In fact, the timing assumption that is usually emphasized is not the only identifying assumption being made in a standard monetary vector autoregression. The timing assumption rules out reverse causality. Output and interest rates are jointly determined. An assumption must be made about whether the contemporaneous correlation between these variables is taken to reflect a causal influence of one on the other or the reverse. This is what the timing assumption does.

Figure 5

Federal Funds Rate Target and 1-Month Eurodollar Rate in 2001 and Early 2002



A: Overview

B: Close-up, detail view

Note: The figure plots the federal funds rate target (the steps) and the one-month Eurodollar rate (smooth line in left panel and the dots in right) at a daily frequency (beginning of day) from December 2001 to March 2002. Dates of changes in the federal funds rate target are indicated in the figure. The dates marked with an asterisk (*) are unscheduled Federal Open Market Committee conference calls. The other dates are scheduled FOMC meetings. The shaded bar in Figure 5B indicates September 11, 2001 (day of the New York terrorist attacks) to September 17 (the day the markets reopened). Figure 5A is very similar to the top panel of Figure 1 in Cochrane and Piazzesi (2002).

But reverse causality is not the only issue when it comes to identifying exogenous variation in policy. Arguably, a much bigger issue is whether monetary policy is reacting to some other piece of information about the current or expected future state of the economy that is not included in the vector autoregression (that is, omitted variables bias). The typical vector autoregression includes a small number of variables and a small number of lags (usually one year worth of lagged values). Any variable not sufficiently well-proxied by these variables is an omitted variable. The omission of these variables leads endogenous variation in policy to be considered exogenous. In the econometrics literature on structural vector autoregression, this omitted variables issue is referred to as the "non-invertibility problem" (Hansen and Sargent 1991; Fernàndez-Villaverde, Rubio-Ramírez, Sargent, and Watson 2007; Plagborg-Møller 2017).

The extremely rich nature of the Fed's information set means that it is arguably hopeless to control individually for each relevant variable. Romer and Romer (2004) propose an interesting alternative approach. Their idea is to control for the Fed's own "Greenbook" forecasts. The idea is that the endogeneity of monetary policy is due to one thing and one thing only: what the Fed thinks will happen to the economy. If one is able to control for this, any residual variation in policy is exogenous. For this reason, the Fed's forecasts are a sufficient statistic for everything that needs to be controlled for.[15]

Romer and Romer's (2004) approach helps answer the question: What is a monetary shock? The Fed does not roll dice. Every movement in the intended federal funds rate is in response to something. Some are in response to developments that directly affect the change in output in the next year. These are endogenous when changes in output over the next year are the outcome variable of interest. But the Fed may also respond to other things: time variation in policymakers' preferences and goals (for example, their distaste for inflation), time variation in policymakers' beliefs about how the economy works, political influences, or pursuit of other objectives (for example, exchange rate stability). Importantly, changes in policy need *not* be unforecastable as long as they are orthogonal to the Fed's forecast of the dependent variable in question. However, changes in policy that are forecastable (for example, forward guidance) are more complicated to analyze since they can have effects both upon announcement and when they are implemented.

What Do We Do with These Shocks?

Any exercise in dynamic causal inference involves two conceptually distinct steps: 1) the construction of the shocks, and 2) the specification used to construct an impulse response once one has the shocks in hand. We have discussed the first of these steps in detail above. But the second step is also important. A specification that imposes minimal structure (apart from linearity) is to directly regress the variable of interest (say, the change in output over the next year) on the shock, perhaps controlling for some variables determined before the shock occurs (pre-treatment controls). This is the specification advocated by Jordà (2005). To construct an

[15] Suppose we are interested in the effect of a change in monetary policy at time t, denoted Δr_t, on the change in output over the next j months, $\Delta^j y_{t+j} = y_{t+j} - y_{t-1}$. The potential concern is that Δr_t may be correlated with some other factor that affects $\Delta^j y_{t+j}$. But this can only be the case if the Fed knows about this other factor, and to the extent that it does, this should be reflected in the Fed's time t forecast of $\Delta^j y_{t+j}$. As Cochrane (2004) emphasizes, controlling for the Fed's time t forecast of $\Delta^j y_{t+j}$ should therefore eliminate all variation in policy that is endogenous to the determination of $\Delta^j y_{t+j}$.

impulse response using this approach, one must run a separate regression for each time horizon that one is interested in plotting.

Standard vector autoregressions construct impulse response functions using a different approach that imposes much more structure: they use the estimated dynamics of the entire vector autoregression system to iterate forward the response of the economy to the shock. This method for constructing impulse responses embeds a new set of quite strong identifying assumptions. In a standard monetary vector autoregression, whether the shocks truly represent exogenous variation in monetary policy only relies on the regression equation for the policy instrument being correctly specified. In contrast, the construction of the impulse response relies on the entire system of equations being a correct representation of the dynamics of all the variables in the system—that is, it relies on the whole model being correctly specified.

It is well-known that the solution to any linear rational expectations model can be represented by a vector autoregression (Blanchard and Kahn 1980; Sims 2002). This idea is the usual defense given regarding the reasonableness of the impulse response construction in a standard vector autoregression. However, to estimate the true vector autoregression, all state variables in the economy must be observable so that they can be included in the system. If this is not the case, the vector autoregression is misspecified and the impulse responses that it yields are potentially biased.[16]

Coibion (2012) has drawn attention to the fact that in Romer and Romer's (2004) results, the peak responses of industrial production and unemployment to a change in the federal funds rate are roughly six times larger than in a standard monetary vector autoregression. In the online appendix to this paper, we revisit this issue by estimating the response of industrial production and the real interest rate to monetary shocks in six different ways. First, we use two different shock series: Romer and Romer's (2004) shock series (as updated and improved by Wieland and Yang 2017) and a shock series from a standard monetary vector autoregression. Second, we estimate the impulse response using three different methods: iterating the vector autoregression dynamics, direct regressions, and the single-equation autoregressive model employed by Romer and Romer (2004). This analysis shows that both the shocks and the method for constructing an impulse response can lead to meaningful differences and help explain the difference in results between Romer and Romer (2004) and standard monetary vector autoregressions. In addition, this analysis shows—as Coibion (2012) emphasizes—that about half of the difference is due to the fact that Romer and Romer's shocks are bigger, that is, they result in larger responses of the real interest rate.

[16] Suppose one of the state variables in the system is not observable. One strategy is to iteratively solve out for that variable. The problem with this is that it typically transforms a vector autogression of order p (VAR(p)) into an infinite order vector autoregression moving average system (VARMA(∞,∞)) in the remaining variables. Thus, the estimation of standard vector autoregressions relies on the assumption that the true infinite-order vector autoregression moving average system, in the variables that the researcher intends to include in the analysis, can be approximated with a vector autoregression of order p. This is a strong assumption that we fear is unlikely to hold in practice. In the online appendix to this paper, we present an example (in the form of a problem set) that illustrates these ideas.

A recent innovation in dynamic causal inference is the use of "external instruments" in vector autoregressions (Stock and Watson 2012; Mertens and Ravn 2013; Stock and Watson 2018). Gertler and Karadi (2015) use this method to estimate the effects of exogenous monetary shocks on output, inflation, and credit spreads. The strength of this method is that it allows researchers to use instrumental variables to identify monetary shocks within the context of a vector autoregression. However, it does not relax the assumptions embedded in using the vector autoregression system to construct the impulse response. We discuss this method in more detail in the online appendix. The use of sign restrictions is another recent development in this area (for example, Uhlig 2017).

Conclusion

Macroeconomics and meteorology are similar in certain ways. First, both fields deal with highly complex general equilibrium systems. Second, both fields have trouble making long-term predictions. For this reason, considering the evolution of meteorology is helpful for understanding the potential upside of our research in macroeconomics. In the olden days, before the advent of modern science, people spent a lot of time praying to the rain gods and doing other crazy things meant to improve the weather. But as our scientific understanding of the weather has improved, people have spent a lot less time praying to the rain gods and a lot more time watching the weather channel.

Policy discussions about macroeconomics today are, unfortunately, highly influenced by ideology. Politicians, policymakers, and even some academics hold strong views about how macroeconomic policy works that are not based on evidence but rather on faith. The only reason why this sorry state of affairs persists is that our evidence regarding the consequences of different macroeconomic policies is still highly imperfect and open to serious criticism. Despite this, we are hopeful regarding the future of our field. We see that solid empirical knowledge about how the economy works at the macroeconomic level is being uncovered at an increasingly rapid rate. Over time, as we amass a better understanding of how the economy works, there will be less and less scope for belief in "rain gods" in macroeconomics and more and more reliance on convincing empirical facts.

■ *We thank Miguel Acosta, Juan Herreño, and Yeji Sung for excellent research assistance. We would like to thank Joshua Angrist, John Cochrane, Gregory Cox, Gauti Eggertsson, Mark Gertler, Adam Guren, Gordon Hanson, Jonathon Hazell, Juan Herreño, Ethan Ilzetzki, Alisdair McKay, Edward Nelson, Serena Ng, Mikkel Plagborg-Møller, Valerie Ramey, David Romer, Timothy Taylor, David Thesmar, Jonathan Vogel, Johannes Wieland, Christian Wolf, and Michael Woodford for valuable comments and discussions. We thank the National Science Foundation (grant SES-1056107) and the Alfred P. Sloan Foundation for financial support.*

References

Acconica, Antonio, Giancarlo Corsetti, and Saverio Simonelli. 2014. "Mafia and Public Spending: Evidence on the Fiscal Multiplier from a Quasi-Experiment." *American Economic Review* 104(7): 2185–2209.

Aguiar, Mark, and Erik Hurst. 2005. "Consumption versus Expenditure." *Journal of Political Economy* 113(5): 919–48.

Angeletos, George-Marios, David Laibson, Andrea Rebetto, Jeremy Tobacman, and Stephen Weinberg. 2001. "The Hyperbolic Consumption Model: Calibration, Simulation, and Empirical Evaluation." *Journal of Economic Perspectives* 15(3): 47–68.

Angrist, Joshua D., Òscar Jordà, and Guido M. Kuersteiner. Forthcoming. "Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited." *Journal of Business and Economic Statistics*.

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

Ashcraft, Adam B. 2005. "Are Banks Really Special? New Evidence from the FDIC-Induced Failure of Healthy Banks." *American Economic Review* 95(5): 1712–30.

Autor David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103(6): 2121–68.

Barro, Robert J., and Charles J. Redlick. 2011. "Macroeconomic Effects from Government Purchases and Taxes." *Quarterly Journal of Economics* 126(1): 51–102.

Basu, Susanto, John G. Fernald, and Miles S. Kimball. 2006. "Are Technology Improvements Contractionary?" *American Economic Review* 96(5): 1418–48.

Beraja, Martin, Erik Hurst, and Juan Ospina. 2016. "The Aggregate Implications of Regional Business Cycles." NBER Working Paper 21956.

Berger, David, Veronica Guerrieri, Guido Lorenzoni, and Joseph Vara. 2018. "House Prices and Consumption Spending." *Review of Economic Studies* 85(3): 1502–42.

Bils, Mark, and Peter J. Klenow. 2004. "Some Evidence on the Importance of Sticky Prices." *Journal of Political Economy* 112(5): 947–85.

Blanchard, Olivier Jean, and Charles M. Kahn. 1980. "The Solution of Linear Difference Models under Rational Expectations." *Econometrica* 48(5): 1305–11.

Blanchard, Olivier, and Roberto Perotti. 2002. "An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output." *Quarterly Journal of Economics* 117(4): 1329–68.

Boppart, Timo, and Per Krusell. 2016. "Labor Supply in the Past, Present, and Future: A Balanced-Growth Perspective." NBER Working Paper 22215.

Burstein, Ariel, Martin Eichenbaum, and Sergio Rebelo. 2005. "Large Devaluations and the Real Exchange Rate." *Journal of Political Economy* 113(4): 742–84.

Calmoris, Charles W., and Joseph R. Mason. 2003. "Consequences of Bank Distress during the Great Depression." *American Economic Review* 93(3): 937–47.

Campbell, Jeffrey R., Charles L. Evans, Jonas D. M. Fisher, and Alejandro Justiniano. 2012. "Macroeconomic Effects of Federal Reserve Forward Guidance." *Brookings Papers on Economic Activity* 43(1): 1–80.

Carvalho, Vasco M., Makoto Nirei, Yukiko U. Saito, and Alirez Tahbaz-Salehi. 2016. "Supply Chain Disruptions: Evidence from the Great East Japan Earthquake." Working Paper 2017-01, Becker Friedman Institution for Research in Economics, University of Chicago.

Catherine, Sylvain, Thomas Chaney, Zongbo Huang, David Alexandre Sraer, and David Thesmar. 2017. "Quantifying Reduced-Form Evidence on Collateral Constraints." Available at SSRN: https://ssrn.com/abstract=2631055.

Chari, V. V., Patrick J. Kehoe, and Ellen R. McGrattan. 2008. "Business Cycle Accounting." *Econometrica* 75(3): 781–836.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge between Structural and Reduced-Form Methods." *Annual Review of Economics* 1: 451–87.

Chetty, Raj. 2012. "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." *Econometrica* 80(3): 969–1018.

Chetty, Raj, Adam Guren, Day Manoli, and Andrea Weber. 2013. "Does Indivisible Labor Explain the Difference between Micro and Macro Elasticities? A Meta-Analysis of Extensive Margin Elasticities." *NBER Macroeconomics Annual* 27(1): 1–56.

Chodorow-Reich, Gabriel. 2014. "The Employment Effects of Credit Market Disruptions: Firm-Level Evidence from the 2008–9 Financial

Crisis." *Quarterly Journal of Economics* 129(1): 1–59.

Chodorow-Reich, Gabriel. 2017. "Geographic Cross-Sectional Fiscal Multipliers: What Have We Learned?" Working Paper, Harvard University OpenScholar.

Chodorow-Reich, Gabriel, John Coglianese, and Loukas Karabarbounis. 2017. "The Limited Macroeconomic Effects of Unemployment Benefit Extensions." Unpublished paper.

Chodorow-Reich, Gabriel, Laura Feiveson, Zachary Liscow, and William Gui Woolston. 2012. "Does State Fiscal Relief during Recessions Increase Employment? Evidence from the American Recovery and Reinvestment Act." *American Economic Journal: Economic Policy* 4(3): 118–45.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *Journal of Political Economy* 113(1): 1–45.

Cochrane, John H. 2004. "Comments on 'A New Measure of Monetary Shocks: Derivation and Implications,' by Christina Romer and David Romer." Presented at NBER EF&G meeting, July 17.

Cochrane, John H., and Monika Piazzesi. 2002. "The Fed and Interest Rates: A High-Frequency Identification." *American Economic Review* 92(2): 90–95.

Coibion, Olivier. 2012. "Are the Effects of Monetary Policy Shocks Big or Small?" *American Economic Journal: Macroeconomics* 4(2): 1–32.

Dupor, Bill, and M. Saif Mehkari. 2016. "The 2009 Recovery Act: Stimulus at the Extensive and Intensive Labor Margins." *European Economic Review* 85: 208–228.

Eggertsson, Gauti B. 2008. "Great Expectations and the End of the Depression." *American Economic Review* 98(4): 1476–1516.

Eggertsson, Gauti B., and Benjamin Pugsley. 2006. "The Mistake of 1937: A General Equilibrium Analysis." *Monetary and Economic Studies* 24(S1): 151–208.

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

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

Faust, Jon, Eric Swanson, and Jonathan H. Wright. 2004. "Do Federal Reserve Policy Surprises Reveal Superior Information about the Economy?" *The BE Journal of Macroeconomics* 4.

Fernández-Villaverde, Jesús, Juan F. Rubio-Ramírez, Thomas J. Sargent, and Mark W. Watson. 2007. "ABCs (and Ds) of Understanding VARs." *American Economic Review* 97(3): 1021–26.

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

Galí, Jordi. 1999. "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations." *American Economic Review* 89(1): 249–71.

Galí, Jordi, David López-Salido, and Javier Vallés. 2007. "Understanding the Effects of Government Spending on Consumption." *Journal of the European Economic Association* 5(1): 227–70.

Gertler, Mark, and Peter Karadi. 2015. "Monetary Policy Surprises, Credit Costs, and Economic Activity." *American Economic Journal: Macroeconomics* 7(1): 44–76.

Goodfriend, Marvin. 2007. "How the World Achieved Consensus on Monetary Policy." *Journal of Economic Perspectives* 21(4): 47–68.

Goodfriend, Marvin, and Robert G. King. 2005. "The Incredible Volcker Disinflation." *Journal of Monetary Economics* 52(5): 981–1015.

Greenstone, Michael, Alesandre Mas, and Hoai-Luu Nguyen. 2017. "Do Credit Market Shocks Affect the Real Economy? Quasi-Experimental Evidence from the Great Recession and 'Normal' Economic Times." Available at SSRN: https://ssrn.com/abstract=2187521.

Hagedorn, Marcus, Iourii Manovskii, and Kurt Mitman. 2015. "The Impact of Unemployment Benefit Extensions on Employment: The 2014 Employment Miracle?" NBER Working Paper 20884.

Hall, Robert E. 2009. "By How Much Does GDP Rise If the Government Buys More Output?" *Brookings Papers on Economic Activity* 40(2): 183–231.

Hansen, Lars Peter, and Thomas J. Sargent. 1991. "Two Difficulties in Interpreting Vector Autoregressions." In *Rational Expectations Econometrics*, edited by L. P. Hansen and T. J. Sargent, 77–119. Boulder, Co.: Westview Press

Hanson, Samuel G., and Jeremy C. Stein. 2015. "Monetary Policy and Long-Term Real Rates." *Journal of Financial Economics* 115(3): 429–48.

Hoover, Keven D., and Stephen J. Perez. 1994. "Post Hoc Ergo Propter Once More: An Evaluation of 'Does Monetary Policy Matter?' in the Spirit of James Tobin." *Journal of Monetary Economics* 34(1): 47–73.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *Quarterly Journal of Economics* 124(4): 1403–48.

Huber, Kilian. 2018. "Disentangling the Effects of a Banking Crisis: Evidence from German Firms and Counties." *American Economic Review* 108(3): 868–98.

Itskhoki, Oleg, and Dmitry Mukhin. 2017. "Exchange Rate Disconnect in General Equilibrium." NBER Working Paper 23401.

Johnson, David S., Jonathan A. Parker, and Nicolas S. Souleles. 2006. "Household Expenditure and the Income Tax Rebates of 2001." *American Economic Review* 96(5): 1589–1610.

Jordà, Òscar. 2005. "Estimation and Inference of Impulse Response by Local Projection." *American Economic Review* 95(1): 161–82.

Kaplan, Greg, and Giovanni L. Violante. 2014. "A Model of the Consumption Response to Fiscal Stimulus Payments." *Econometrica* 82: 1199–1239.

King, Robert G., Charles I. Plosser, and Sergio T. Rebelo. 1988. "Production, Growth and Business Cycles: I. The Basic Neoclassical Model." *Journal of Monetary Economics* 21(2–3): 195–232.

King, Robert G., and Sergio T. Rebelo. 1999. "Resuscitating Real Business Cycles." In *Handbook of Macroeconomics*, edited by John B. Taylor and Michael Woodford, 927–1007. Amsterdam, Holland: Elsevier.

Klenow, Peter J., and Oleksiy Kryvtsov. 2008. "State-Dependent or Time-Dependent Pricing: Does It Matter for Recent U.S. Inflation." *Quarterly Journal of Economics* 123: 863–904.

Kydland, Finn E., and Edward C. Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50(6): 1345–70.

Leeper, Eric M. 1997. "Narrative and VAR Approaches to Monetary Policy: Common Identification Problems." *Journal of Monetary Economics* 40(3): 641–57.

Martin, Pilippe, and Thomas Philippon. 2017. "Inspecting the Mechanism: Leverage and the Great Recession in the Eurozone." *American Economic Review* 107(7): 1904–37.

Meese, Richard A., and Kenneth Rogoff. 1983. "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?" *Journal of International Economics* 14(1–2): 3–24.

Mehra, Rajnish, and Edward C. Prescott. 1985. "The Equity Premium: A Puzzle." *Journal of Monetary Economics* 15(2): 145–61.

Mertens, Karel, and Morten O. Ravn. 2012. "Empirical Evidence on the Aggregate Effects of Anticipated and Unanticipated US Tax Policy Shocks." *American Economic Journal: Economic Policy* 4(2): 145–81.

Mertens, Karel, and Morten O. Ravn. 2013. "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States." *American Economic Review* 103(4): 1212–47.

Mian, Ataf, Kamalesh Rao, and Amir Sufi. 2013. "Household Balance Sheets, Consumption, and the Economic Slump." *Quarterly Journal of Economics* 128(4): 1687–1726.

Mian, Atif, and Amir Sufi. 2014. "What Explains the 2007–2009 Drop in Employment?" *Econometrica* 82(6): 2197–2223.

Miyamoto, Wataru, Thuy Lan Nguyen, and Dmitry Sergeyev. 2018. "Government Spending Multipliers under the Zero Lower Bound: Evidence from Japan." *American Economic Journal: Macroeconomics* 10(3): 247–77.

Mondragon, John. 2018. "Household Credit and Employment in the Great Recession." Nielsen Dataset Paper 1-025, Kilts Center for Marketing at Chicago Booth. Available at SSRN: https://ssrn.com/abstract=2521177.

Mussa, Michael. 1986. "Nominal Exchange Rate Regimes and the Behavior of Real Exchange Rates: Evidence and Implications." *Carnegie-Rochester Conference Series on Public Policy* 25: 117–214.

Nakamura, Emi, and Jón Steinsson. 2008. "Five Facts about Prices: A Reevaluation of Menu Cost Models." *Quarterly Journal of Economics* 123(4): 1415–64.

Nakamura, Emi, and Jón Steinsson. 2014. "Fiscal Stimulus in a Monetary Union: Evidence from US Regions." *American Economic Review* 104(3): 753–92.

Nakamura, Emi, and Jón Steinsson. Forthcoming. "High Frequency Identification of Monetary Non-Neutrality: The Information Effect." *Quarterly Journal of Economics*.

Nekarda, Christopher J., and Valerie A. Ramey. 2011. "Industry Evidence on the Effects of Government Spending." *American Economic Journal: Macroeconomics* 3(1): 36–59.

Nelson, Edward. 2005. "The Great Inflation of the Seventies: What Really Happened?" *The B.E. Journal of Macroeconomics* 5(1).

Parker, Jonathan A., Nicholas S. Souleles, David S. Johnson, and Robert McClelland. 2013. "Consumer Spending and the Economic Stimulus Payments of 2008." *American Economic Review* 103(6): 2530–53.

Peek, Joe, and Eric S. Rosengren. 2000. "Collateral Damage: Effects of the Japanese Bank Crisis on Real Activity in the United States." *American Economic Review* 90(1): 30–45.

Plagborg-Møller, Mikkel. 2017. "Bayesian Inference on Structural Impulse Response Functions." Unpublished paper, Princeton University.

Prescott, Edward C. 1986. "Theory Ahead of Business-Cycle Measurement." *Carnegie-Rochester Conference Series on Public Policy* 25: 11–44.

Ramey, Valerie A. 2011. "Identifying Government Spending Shocks: It's All in the Timing." *Quarterly Journal of Economics* 126(1): 1–50.

Ramey, Valerie A. 2016. "Macroeconomic Shocks and Their Propagation." Chap. 3 of *Handbook of Macroeconomics,* vol. 2A. edited by John B.

Taylor and Harald Uhlig, 71–162. Amsterdam, Holland: Elsevier.

Ramey, Valerie A., and Sarah Zubairy. 2018. "Government Spending Multipliers in Good Times and in Bad: Evidence from U.S. Historical Data." *Journal of Political Economy* 126(2): 850–901.

Romer, Christina D., and David H. Romer. 1989. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." In *NBER Macroeconomics Annual*, edited by O. J. Blanchard and S. Fischer, 121–170. Cambridge, MA: MIT Press.

Romer, Christina D., and David H. Romer. 1994. "Monetary Policy Matters." *Journal of Monetary Economics* 34(1): 75–88.

Romer, Christina D., and David H. Romer. 1997. "Identification and the Narrative Approach: A Reply to Leeper." *Journal of Monetary Economics* 40(3): 659–65.

Romer, Christina D., and David H. Romer. 2000. "Federal Reserve Information and the Behavior of Interest Rates." *American Economic Review* 90(3): 429–57.

Romer, Christina D., and David H. Romer. 2002. "The Evolution of Economic Understanding and Postwar Stabilization Policy." In *Rethinking Stabilization Policy*, 11–78. Kansas City: Federal Reserve Bank of Kansas City.

Romer, Christina D., and David H. Romer. 2004. "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review* 94(4): 1055–84.

Rotemberg, Julio J., and Michael Woodford. 1997. "An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy." In *NBER Macroeconomics Annual 1997*, edited by B. S. Bernanke and J. J. Rotemberg, 297–346. Cambridge, MA: MIT Press.

Rudebusch, Glenn D. 1998. "Do Measures of Monetary Policy in a VAR Make Sense?" *International Economic Review* 39(4): 907–931.

Saiz, Albert. 2010. "The Geographic Determinants of Housing Supply." *Quarterly Journal of Economics* 125(3): 1253–96.

Shapiro, Michael D. 1994. "Federal Reserve Policy: Cause and Effect." In *Monetary Policy*, edited by N. G. Mankiw, 307–334. University of Chicago Press.

Shimer, Robert. 2005. "The Cyclical Behavior of Equilibrium Unemployment and Vacancies." *American Economic Review* 95(1): 25–49.

Shimer, Robert. 2009. "Convergence in Macroeconomics: The Labor Wedge." *American Economic Journal: Macroeconomics* 1(1): 280–97.

Shoag, Daniel. 2015. "The Impact of Government Spending Shocks: Evidence on the Multiplier from State Pension Plan Returns." Working Paper, Harvard University.

Sims, Christopher. 2002. "Solving Linear Rational Expectations Models." *Journal of Computational Economics* 20(1–2): 1–20.

Sinai, Todd, and Nicholas S. Souleles. 2005. "Owner-Occupied Housing as a Hedge against Rent Risk." *Quarterly Journal of Economics* 120(2): 763–89.

Smith, Anthony A. 2008. "Indirect Inference." In *The New Palgrave Dictionary of Economics*, Second Edition, edited by S. Durlauf and L. E. Blume. London: Palgrave Macmillan. Also available at: http://www.econ.yale.edu/smith/palgrave7.pdf.

Stock, James H., and Mark W. Watson. 2012. "Disentangling the Channels of the 2007-09 Recession." *Brookings Papers on Economic Activity* 43(1): 81–135.

Stock, James H., and Mark W. Watson. 2018. "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments." NBER Working Paper 24216.

Suárez Serrato, Juan Carlos, and Philippe Wingender. 2016. "Estimating Local Fiscal Multipliers." NBER Working paper 22455.

Uhlig, Harald. 2017. "Shocks, Sign Restrictions, and Identification." Chap. 4 in *Advances in Economics and Econometrics: Eleventh World Congress*, vol. 2, edited by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, 97–127. Cambridge University Press.

Velde, François R. 2009. "Chronicle of a Deflation Unforetold." *Journal of Political Economy* 117(4): 591–634.

Wieland Johannes F., and Mu-Jeung Yang. 2017. "Financial Dampening." Unpublished paper, University of California, San Diego.

Wilson, Daniel J. 2012. "Fiscal Spending Jobs Multipliers: Evidence from the 2009 American Recovery and Reinvestment Act." *American Economic Journal: Economic Policy* 4(3): 251–82.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 87–112

The State of New Keynesian Economics: A Partial Assessment

Jordi Galí

In August 2007, when the first signs emerged of what would come to be the most damaging global financial crisis since the Great Depression, the New Keynesian paradigm was dominant in macroeconomics. It was taught in economics programs all over the world as the framework of reference for understanding fluctuations in economic activity and inflation and their relation to monetary and fiscal policies. It was widely adopted by researchers as a baseline model that could be used flexibly to analyze a variety of macroeconomic phenomena. The New Keynesian model was also at the core of the medium-scale, dynamic stochastic general equilibrium (DSGE) models developed and used by central banks and policy institutions throughout the world.

Ten years later, tons of ammunition has been fired against modern macroeconomics in general, and against dynamic stochastic general equilibrium models that build on the New Keynesian framework in particular. The criticisms have focused on the failure of these models to predict the crisis, a weakness often attributed to their lack of a financial block in the model that could account for the key factors behind the crisis, whose origin was largely financial.[1] Other aspects of the New

[1] See, for example, Lindé, Smets, and Wouters (2016) for an evaluation of the empirical performance of a standard medium-scale DSGE model during the financial crisis and its aftermath, as well as a discussion of the kind of changes needed to improve that performance. Del Negro and Schorfheide (2013) evaluate the forecasting performance of a similar DSGE model during the Great Recession, and conclude that a

■ *Jordi Galí is a Senior Researcher, Centre de Recerca en Economia Internacional (CREi); Professor of Economics, Universitat Pompeu Fabra; and Research Professor, Barcelona Graduate School of Economics, all in Barcelona, Spain. His email address is jgali@crei.cat.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.87 doi=10.1257/jep.32.3.87

Keynesian model and its extensions that have been the target of criticism include the assumptions of rational expectations, perfect information, and an infinitely-lived representative household.

Those criticisms notwithstanding, the New Keynesian model arguably remains the dominant framework in the classroom, in academic research, and in policy modeling. In fact, one can argue that over the past ten years the scope of New Keynesian economics has kept widening, by encompassing a growing number of phenomena that are analyzed using its basic framework, as well as by addressing some of the criticisms raised against it. Much recent research, for instance, has been devoted to extending the basic model to incorporate financial frictions (as described by Gertler and Gilchrist in this issue). In addition, the New Keynesian model has been the framework of choice in much of the work aimed at evaluating alternative proposals to stimulate the economy in the face of the unusual circumstances triggered by the crisis, including the use of fiscal policy and unconventional monetary policies.[2]

The present paper takes stock of the state of New Keynesian economics by reviewing some of its main insights and by providing an overview of some recent developments. In particular, I discuss some recent work on two very active research programs: the implications of the zero lower bound on nominal interest rates and the interaction of monetary policy and household heterogeneity. Finally, I discuss what I view as some of the main shortcomings of the New Keynesian model and possible areas for future research.

The New Keynesian Model: A Refresher

Modern New Keynesian economics can be interpreted as an effort to combine the methodological tools developed by real business cycle theory with some of the central tenets of Keynesian economics tracing back to Keynes's own *General Theory*, published in 1936.

The hallmark of the approach to modeling economic fluctuations pioneered by real business cycle theorists is a reliance on *dynamic, stochastic, general equilibrium* frameworks. At some level, these terms describe what seem natural features of any model that seeks to explain economic fluctuations, and as such, these features have been fully adopted by New Keynesian economics. (To put it differently: It is easy to imagine the criticisms that modern macro would receive if it relied on models that were static rather than dynamic, deterministic rather than stochastic, and partial

version of that model augmented with financial frictions and external information compares well with Blue Chip consensus forecasts, especially over the medium and long run.

[2] See, for example, Blanchard, Erceg, and Lindé (2016) for an analysis of the effectiveness of fiscal policy to stimulate the recovery of the euro area economy using a DSGE model as a framework of reference. Del Negro, Eggertsson, Ferrero, and Kiyotaki (2017) use a standard DSGE model augmented with liquidity frictions to evaluate some of the quantitative easing policies undertaken by the Fed in the wake of the financial crisis.

rather than general equilibrium!) In practice, the real business cycle approach takes the form of a set of equations that describe, in a highly aggregative manner: 1) the behavior of households, firms, and policymakers, 2) some market clearing and/or resource constraints, and 3) the evolution of one or more exogenous variables that are the ultimate source of fluctuations in the economy. More controversial may be the assumption, widely found in both real business cycle and New Keynesian models, that the behavior of households and firms (and, in some instances, of policymakers as well) is the outcome of an *optimization problem*, solved under the assumption of *rational expectations* (though a strand of the recent literature, not reviewed here, has examined the consequences of relaxing the latter assumption).

What does New Keynesian economics add to the standard real business cycle apparatus? One can pinpoint three significant modifications. First, it introduces nominal variables explicitly: prices, wages, and a nominal interest rate. Second, it departs from the assumption of perfect competition in the goods market, allowing for positive price markups. Third, it introduces nominal rigidities, generally using the formalism proposed by Calvo (1983), whereby only a constant fraction of firms, drawn randomly from the population, are allowed to adjust the price of their good. The assumption of imperfect competition is often extended to the labor market as well, with the introduction of wage rigidities (nominal or real).

The resulting framework has two key properties. Exogenous changes in monetary policy have nontrivial effects on real variables, not only on nominal ones. In addition, and more importantly, the economy's equilibrium response to *any* shock is not independent of the monetary policy rule in place, thus opening the door to a meaningful analysis of alternative monetary policy rules.

To build some intuition for how this framework leads to a breakdown of monetary policy neutrality, it is useful to lay out a simple version of the New Keynesian model (with sticky prices but flexible wages). It is composed of three relationships.

First, the *dynamic IS equation* (named after the IS curve in the celebrated IS-LM model) states that the current output gap is equal to the difference between the expected output gap one period in the future and an amount that is proportional to the gap between the real interest rate and the natural rate of interest. The "output gap" is the difference between output and the potential or "natural" output. Natural output and the natural rate of interest are the values that those variables would take in equilibrium if prices were fully flexible. In algebraic terms, the relationship is

$$\tilde{y}_t = E_t\{\tilde{y}_{t+1}\} - \sigma^{-1}(i_t - E_t\{\pi_{t+1}\} - r_t^n)$$

where $\tilde{y}_t$ is the output gap (given by the difference between log output y_t and log natural output y_t^n), i_t is the nominal rate, π_t denotes inflation, and r_t^n is the natural rate of interest.

Second, the *New Keynesian Phillips curve* states that inflation depends on expected inflation one period ahead and the output gap.[3] Thus, it adds an expectation term to the conventional Phillips curve, and can be written out as:

$$\pi_t = \beta E_t\{\pi_{t+1}\} + \kappa \tilde{y}_t.$$

The third relationship is an *interest rate rule*, which describes how the nominal rate of interest is determined. This condition is typically linked to the conduct of monetary policy. Thus, an interest rate rule frequently used in the literature as an approximation to the conduct of monetary policy in advanced economies (at least in normal times) is a Taylor-type rule in which nominal interest rates traditionally rise and fall based on the current inflation rate and detrended output (for example, Taylor 1993), but in which monetary policy at a given time can be tighter or looser than the historical pattern. This relationship can be written as

$$i_t = \phi_\pi \pi_t + \phi_y \hat{y}_t + v_t,$$

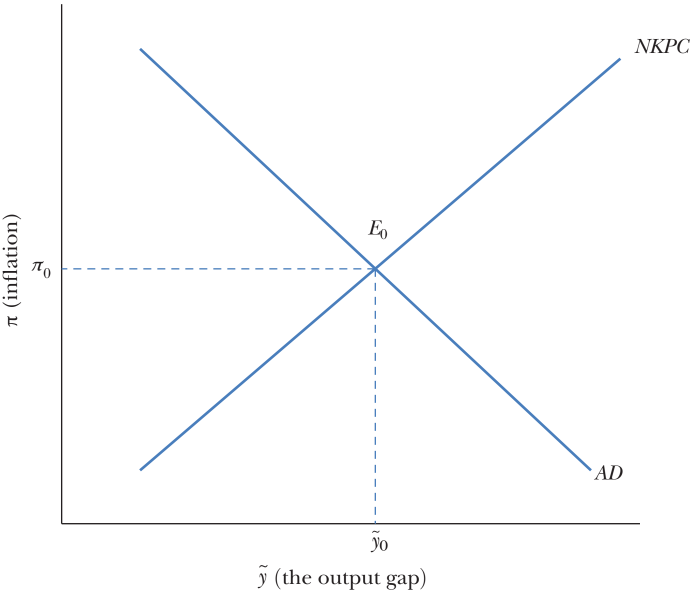
where $\hat{y}_t$ denotes the log deviation of output from steady state, and v_t is an exogenous monetary policy shifter following some stochastic process.

Figure 1 represents the equilibrium of the above economy. The AD schedule (after "aggregate demand") combines the dynamic IS equation and the interest rate rule, giving rise to an inverse relation between inflation and the output gap, for any given expectations. The NKPC (New Keynesian Phillips Curve) schedule represents a positive relation between the same two variables implied by the New Keynesian Phillips curve, given inflation expectations. The economy's equilibrium is determined by the intersection of the two schedules (point E_0).

A New Keynesian model based on these three relationships yields several interesting insights. As noted earlier, the model implies that monetary policy is not neutral. In particular, this non-neutrality has (at least) two dimensions. First, an exogenous monetary policy shock will affect not only nominal variables, but also real ones (like output). In particular, an exogenous tightening of monetary policy (that is, a persistent increase in v_t) raises both nominal and real rates, leading to a fall in output and inflation, while leaving the natural rates unchanged. In Galí (2015), I discuss the implied response of a calibrated New Keynesian model to different types

[3] See Woodford (2003) or Galí (2015) for a detailed derivation of these first two equations and a discussion of the underlying assumptions. The first equation can be derived by combining the Euler equation describing the optimal consumption behavior of the representative household with a goods market clearing condition requiring that output must equal consumption. The second equation can be derived in two stages. In a first stage, a relation between inflation, expected inflation and the markup gap (that is, the log deviation of average markup from the desired markup) can be derived by aggregating the optimal price setting decisions of firms subject to constraints on the frequency with which they can adjust prices. That relation is combined with a labor supply equation, a goods market clearing condition, and an aggregate production function to obtain a simple relation linking the markup gap to the output gap, thus giving rise to the relationship in the text.

Figure 1
The Basic New Keynesian Model



Note: The AD schedule (after "aggregate demand") combines the dynamic IS equation and the interest rate rule, giving rise to an inverse relation between inflation and the output gap, for any given expectations. The NKPC schedule represents the positive relation between the same two variables implied by the New Keynesian Phillips Curve, given inflation expectations. The economy's equilibrium is determined by the intersection of the two schedules (point E_0).

of shocks. In particular, the model's predictions regarding the effects of monetary policy shocks are in line (at least qualitatively), with much of the empirical evidence on the effects of those shocks, as found among others in Christiano, Eichenbaum, and Evans (1999).

Second, monetary non-neutrality in this context also arises because the response of output (and other real variables) to a nonmonetary shock—that is, a shock that changes the natural levels of output y_t^n and/or interest rate r_t^n—is not invariant to the monetary policy rule adopted by the central bank. Interestingly, when the interest rate rule shown above is calibrated in a way consistent with US evidence for the post–1982 period, the model implies responses to technology shocks consistent with the empirical evidence, including a countercyclical response of employment (for example, see Galí 1999; Basu, Fernald, and Kimball 2006).

The New Keynesian model also generates normative insights for the conduct of monetary policy. One finding is that if the central bank applies a rule that adjusts the policy interest rate sufficiently strongly in response to variations in inflation and output (a condition known as the Taylor principle), then the economy will have a unique equilibrium.[4] Otherwise, the equilibrium is locally indeterminate, opening the door to fluctuations driven by self-fulfilling revisions in expectations (sometimes known as "sunspot fluctuations"). Clarida, Galí, and Gertler (2000) provide evidence suggesting that the local uniqueness condition may not have been satisfied during the pre-Volcker era, potentially giving rise to unnecessary instability and providing an explanation for the macroeconomic turbulence of that period.

Beyond simple rules like the Taylor-type rule described above, the literature has sought to characterize the optimal monetary policy, defined as the policy that maximizes welfare for the representative household. One useful formulation suggests that the optimal monetary policy should consider three sources of welfare losses: 1) fluctuations in the gap between output and its efficient level (the so-called "welfare-relevant output gap"); 2) fluctuations in inflation, which generate losses due to the misallocation of resources caused by the associated price dispersion; and 3) an average (steady state) level of output which is itself inefficiently low, due to uncorrected real distortions (as one example, arising from monopolistic competition).

In the *special case* in which the natural level of output corresponds to the efficient level of output at all times, then welfare losses result only from fluctuations in the output gap, $\tilde{y}$, and fluctuations in inflation, π_t. The optimal policy in that special case requires that inflation be fully stabilized at zero. Notice that the New Keynesian Phillips curve implies that such a strict inflation targeting policy has an important byproduct: it stabilizes the output gap at zero, thus making output equal to its natural (and, by assumption, efficient) level. This property is sometimes referred to as the Divine Coincidence (for discussion, see Blanchard and Galí 2007). As a result, welfare losses in this setting will be zero and the economy attains its first-best allocation.

However, the previous extreme result holds only when the flexible price (or natural) equilibrium allocation is optimal—that is, when nominal rigidities are the only distortion in the economy. More generally, the presence of real frictions is likely to drive a wedge between the natural and efficient levels of output. As a result, the steady state itself may be inefficient, or the presence of real frictions may imply an inefficient response of natural output to some shocks, or both. As a result, a trade-off emerges between price stability and the attainment of an efficient level of economic activity, thus giving rise to a nontrivial optimal monetary policy problem. It turns out that the optimal policy—along with its associated output gap and inflation outcomes—depends on the assumptions regarding the extent to which the central bank can credibly commit to a state-contingent plan. Standard treatments of the optimal monetary policy problem and its consequences have focused on the extreme cases of full discretion (period-by-period re-optimization) and full commitment (a once-and-for-all choice of an optimal plan, which is subsequently followed

[4] As shown by Bullard and Mitra (2002), the required condition takes the form $\kappa(\phi_\pi - 1) + (1 - \beta)\phi_y > 0$.

through even if the policymaker may be tempted to renege from it, the so-called "time-inconsistency problem").[5]

The study of the optimal interest rate policy in the context of the New Keynesian model has yielded several interesting insights, and in particular about the nature of the gains from commitment and the kind of inefficient outcomes or biases implied by discretionary policies. For example, the presence of an inefficiently low steady state output, combined with the lack of commitment, generates a (suboptimal) positive *inflation bias*, similar to that uncovered by Kydland and Prescott (1980) and Barro and Gordon (1983) in the context of an earlier generation of monetary models with non-neutralities. Most interestingly, even when the steady state is efficient, gains from commitment arise in the presence of shocks that imply an inefficient response of natural output, as would arise in the presence of certain real imperfections. Those gains result from the ability of a central bank with commitment to influence expectations of future inflation and output gaps, which makes it possible to smooth over time the deviations from the first-best allocation, thus reducing the implied losses. By contrast, in the absence of commitment, the central bank has to rely exclusively on its ability to affect the *current* output gap, which leads to excessive fluctuations in both inflation and the gap between output and its efficient level, and hence to larger welfare losses. The resulting excess volatility associated with the discretionary policy is sometimes referred to as *stabilization bias,* and it may coexist with an optimal *average* level of inflation (in contrast with the case of an inflation bias).

While the notion of a once-and-for-all commitment to an optimal state-contingent monetary policy plan is of course unrealistic as a practical policy strategy, the analysis of the optimal policy under commitment establishes a useful benchmark that can be used to inform the search for simpler rules that can approximate such a policy. Specifically, the analysis of the properties of the equilibrium under the optimal monetary policy with commitment often seem to imply a stationary price level, which in turn provides a possible rationale for the adoption of a price-level–targeting interest rate rule (as one example, see Vestin 2006).

Many other interesting insights regarding the optimal design of monetary policy have emerged from the analyses of relatively straightforward extensions of the basic New Keynesian model described above. A selection of examples of such extensions include allowances for staggered wage setting (Erceg, Henderson, and Levin 2000), some backward-looking price setting (Steinsson 2003), open economy considerations (Clarida, Galí, and Gertler 2002; Galí and Monacelli 2005), deviations from rational expectations (Evans and Honkapohja 2003; Woodford 2010), labor market frictions (Trigari 2009; Blanchard and Galí 2010), uncertainty shocks (Basu and Bundick 2017), and others. The next two sections focus on two specific extensions of the basic New Keynesian model that have drawn considerable attention

[5] See Clarida, Galí, and Gertler (1999) for an analysis and discussion of the resulting optimal monetary policy problem under discretion and under commitment. For an analysis of some intermediate cases, see Schamburg and Tambalotti (2007) and Debortoli and Lakdawala (2016).

in recent years and triggered a good amount of research: the zero lower bound on the nominal interest rate and the heterogeneity of households.

The Zero Lower Bound

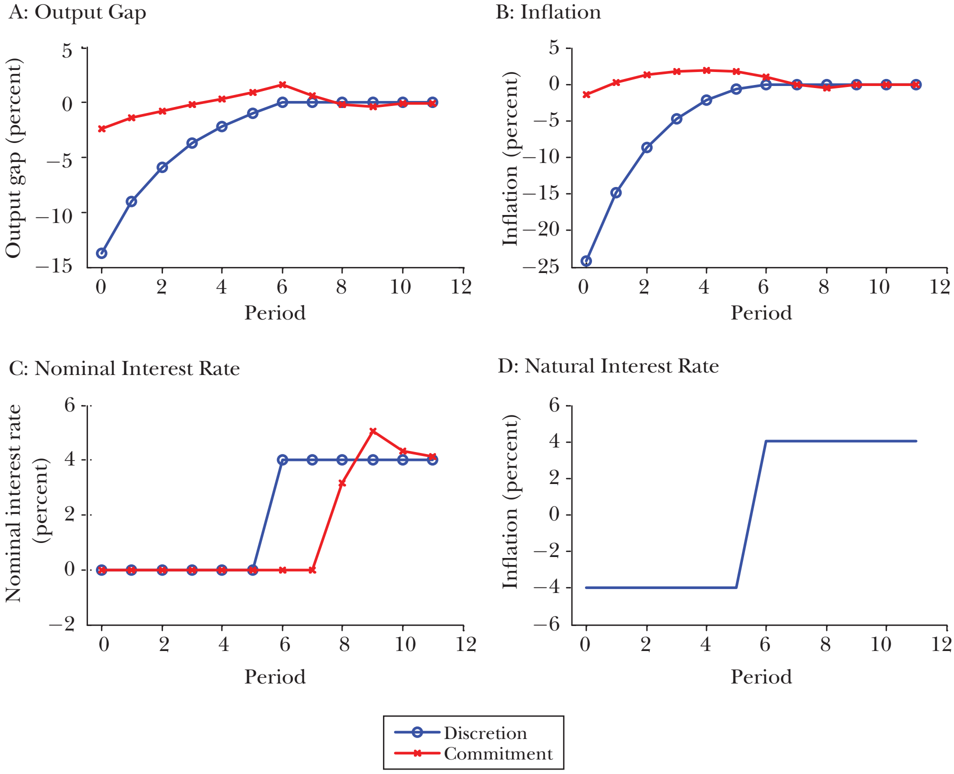
The possibility of a nimble response of central banks to the recessionary and deflationary forces triggered by the financial crisis was seemingly jeopardized when, after being successively reduced, policy rates attained the lower bound of (nearly) zero percent. The basic New Keynesian model, described in the previous section, ignores the existence of the zero lower bound. However, a number of papers, originally motivated by the Japanese experience with a liquidity trap starting in the 1990s, adopted the New Keynesian framework to analyze the implications of a binding zero lower bound.

To illustrate some of the insights of that analysis, let us consider the case of an efficient natural equilibrium (that is, the gap between the efficient and the natural rate of output is zero). In the absence of the zero lower bound constraint, the optimal policy implies full stabilization of the output gap and inflation, as discussed above.

Now consider an economy that is at a zero-inflation, zero-output-gap steady state. Then a one-off episode occurs, with a temporary, but persistent adverse shock to the natural rate of interest, r_t^n, that brings that variable into negative territory. With a zero lower bound constraint, there is an inability to match the drop in the natural rate of interest with a commensurate reduction in the policy interest rate. Using the dynamic IS relationship shown earlier, the result of a nominal interest rate stuck above its natural rate will generate a persistent negative output gap (given the initial zero inflation). In turn, the New Keynesian Phillips curve relationship shows that a negative output gap will be a source of persistent deflation. Indeed, this leads to a higher real interest rate and, thus, to an even larger gap between that variable and its natural counterpart, deepening further the initial recession. An analysis of the optimal design of monetary policy in the presence of a zero lower bound on the nominal rate closely related to this description can be found in Jung, Teranishi, and Watanabe (2005) and Eggertsson and Woodford (2003). Both papers study the case of a fully unanticipated, once-and-for-all adverse shock to the natural rates, which pushes the optimizing central bank against the zero lower bound.

Figure 2, based on the analysis in Galí (2015), illustrates some of the implications of the zero lower bound for the conduct of monetary policy. It simulates the response to an unanticipated negative demand shock that lowers the natural rate of interest from its normal steady state level of 4 percent to −4 percent (both in annual terms) between periods 0 through 5 (see panel D). In period 6, the natural rate returns to its initial value, something that is assumed to be (correctly) anticipated as of period 0, when the shock hits. In the absence of a zero lower bound, price stability and a zero output gap could be maintained in the face of the adverse disturbance if only the central bank were to lower the interest rate to −4 percent for the duration of the shock, thus tracking the path of the

Figure 2
Discretion versus Commitment in the Presence of the Zero Lower Bound



natural rate. However, the existence of the zero lower bound makes that option unfeasible.

Given this setting, the nature of the optimal interest rate policy depends on the extent to which the central bank can commit to future actions. The line with circles plots the response of the output gap, inflation, and the nominal interest rate (panels A, B, and C) under the optimal discretionary policy—that is, a policy without commitment. In response to the adverse shock, the central bank lowers the nominal rate to zero and keeps it there until the shock goes away, and then returns the interest rate to its initial level of 4 percent, consistent with price stability. Both output and inflation experience large declines in response to the shock and take persistent negative values until the adverse disturbance vanishes, at which point the central bank can fully restore price stability and close the output gap.

Also in the first three panels, the line with crosses displays the equilibrium responses under the optimal policy with commitment. In this case, the central bank credibly promises that it will keep the nominal rate at zero even after the shock is no longer effective (in this simulated example, for two periods longer). That policy leads to a small deviation from zero inflation and zero output gap in subsequent periods, implying a welfare loss relative to that first-best outcome. But that loss is more than offset by the gains resulting from the much greater stability in earlier periods, when the disturbance is active. That optimal policy with commitment can be interpreted as an illustration of the power of *forward guidance* policies, which are policies that aim at influencing current macro outcomes through the management of expectations about future policy settings. Such policies have been openly adopted by central banks like the European Central Bank and the Federal Reserve in the aftermath of the Great Recession, in the face of the slow recovery.[6]

The previous example illustrates the monetary policy implications of a fully unanticipated, one-off temporary drop in the natural rate of interest to a negative level. A number of authors have instead analyzed an economy where the natural rate of interest is subject to recurrent shocks. In those economies, the possibility of hitting the zero lower bound constraint in the future affects how the economy responds to shocks (and to policy) even when the zero lower bound is not binding. Adam and Billi (2006, 2007) and Nakov (2008) study the implications of the zero lower bound for the optimal design of monetary policy in a stochastic setting, with and without commitment, when that constraint is occasionally (but recurrently) binding.

Several insights emerge from that line of analysis. First, the optimal policy implies a nonlinear response to shocks, with the central bank reducing nominal rates more aggressively in response to adverse shocks, in order to reduce the probability of a binding zero lower bound down the road and to counteract the adverse effects of that possibility (and their anticipation) on aggregate demand. Second, under commitment, the optimal policy calls for sustained monetary easing even when the natural rate is no longer negative. Third, the gains from commitment (relative to discretion) are much larger when the possibility of a zero lower bound exists than in the absence of such a constraint. Finally, as stressed by Nakov (2008), a large fraction of the gains from commitment can be reaped by adopting a price-level targeting rule. Because this rule targets the level of prices, rather than the inflationary change in price level, it calls for a period of "catching up" after inflation has been below its target level for a time—not just a return to the target level. Such a rule also reduces the incidence of a binding zero lower bound considerably.

Rogoff (1985) made a case for appointing a "conservative" central banker (that is, one that puts more weight than society on inflation stabilization), in the presence of a conventional inflation bias. Nakata and Schmidt (2016) provide a new rationale, connected to the zero lower bound, for such a policy, even in the absence of an inflation bias. They show that, under an optimal discretionary policy, the anticipation of an occasionally binding zero lower bound implies that on average inflation falls below target and the output gap is positive, even when the zero lower bound is not binding. Delegating monetary policy to a "conservative" central banker is generally desirable since the latter will keep inflation closer to target (at the cost of an even larger output gap) when the zero lower bound is not binding, with the

[6] For example, Woodford (2013) discusses the forward guidance policies implemented by different central banks and their connection with the theoretical analyses in the literature.

anticipation of that policy providing a highly welcome additional stimulus when the zero lower bound is binding, and improving social welfare.

The Forward Guidance Puzzle

The forward guidance puzzle can be stated as follows: In the context of the basic New Keynesian model, and under the assumption of rigid prices, the effect on output of an anticipated change in the policy rate of a given size and duration is independent of the timing of its implementation. In other words, the effects of a temporary 1 percent increase in the policy rate 100 years from now is predicted in the basic New Keynesian model to be the same as if the increase were to take place immediately or in the near future. This forward guidance puzzle was first discussed by Carlstrom, Fuerst, and Paustian (2015) and Del Negro, Giannoni, and Patterson (2012).

The reason behind that prediction is that the dynamic IS relationship presented earlier implies no discounting of the expected output gap and, hence, no discounting of future interest rates. To see this, iterate that dynamic IS relationship forward, noting that the expected output gap in the next period depends on the expected interest gap one period ahead and the expected output gap in following period, and so on. Moreover, assume that the output gap is expected to converge to zero asymptotically, and that the price level is rigid (with inflation equal to zero), then the forward guidance puzzle arises: the current output gap depends on the sum of current and future interest rates, all of them having the same weight.

The puzzle is amplified if we relax the assumption of fully rigid prices and let inflation be determined by a New Keynesian Phillips curve relationship. In that case, the farther is the horizon of implementation of a given change in the policy interest rate, the longer are its effects on output, and hence the larger and more persistent is the response of inflation. For any given path of the nominal interest rate, the persistent effect of inflation works in the direction of changing the real interest rate in a way that further amplifies the effects on output and inflation—leading to a strong nonlinear effect due to the accumulation of feedback effects.

Several authors have sought to address the forward guidance puzzle with modifications to the benchmark New Keynesian model. Typically, such modifications lead to some kind of discounting by households. Examples of such modifications include the introduction of finite lives (Del Negro, Giannoni, and Patterson 2012), incomplete markets with bounded rationality (Farhi and Werning 2017) or without bounded rationality (McKay, Nakamura, and Steinsson 2016, 2017), lack of common knowledge (Angeletos and Lian forthcoming), and behavioral discounting (Gabaix 2017).

With such modifications, the effects of anticipated changes in the policy interest rate on current output do decline with the horizon of implementation, given the path of inflation. However, once inflation is allowed to respond, the presence of discounting reduces the effect on the output gap or inflation of any anticipated change in the real interest rate, but it does not overturn the prediction that the size of such an effect increases with the horizon of implementation.

Self-Fulfilling Deflation Traps and the Zero Lower Bound

Much of the analysis based on the New Keynesian model has a local nature: specifically, it is carried out using a linear approximation to the equilibrium conditions around a steady state consistent with the inflation target (which is typically zero). By construction, that analysis limits our understanding of the economy's behavior far from the assumed steady state. Several papers have explored the properties of equilibria of the New Keynesian model from a global perspective.

The work of Benhabib, Schmitt-Grohé, and Uribe (2001) triggered much of the research on this front. They showed that a Taylor-type interest rate rule satisfying the zero lower bound constraint and consistent with a locally unique equilibrium around the steady state associated with the targeted inflation rate necessarily implies the existence of another steady state. They label this other steady state as a "liquidity trap," in which the interest rate is zero or near-zero and inflation is below the targeted level and possibly negative. Furthermore, and more worrisome, they showed that an infinite number of equilibrium trajectories exist that converge to the liquidity trap steady state. Accordingly, a central bank's adoption of a Taylor rule is not a guarantee of stability, even if the rule satisfies the conditions for a locally unique equilibrium. In a companion paper, Benhabib, Schmitt-Grohé, and Uribe (2002) propose a set of alternative monetary and fiscal rules that can be activated when the economy enters a path leading to the liquidity trap steady state. For example, in one case the proposed rule features a strong fiscal stimulus in the form of lower taxes; in another, a switch to a rule that would peg the rate of money growth. Under those rules, any path converging to the liquidity trap would violate the intertemporal budget constraints of the government and households, and can thus be ruled out as an equilibrium path.[7]

Several papers have provided "quantitative" applications of the multiplicity of global equilibrium implied by the zero lower bound in the New Keynesian model. Schmitt-Grohé and Uribe (2017), Aruoba, Cuba-Borda, and Schorfheide (2018), and Jarociński and Maćkowiak (2018) use a quantitative New Keynesian model with global multiplicity to interpret the prolonged recession and persistently low inflation in many advanced economies in the wake of their financial crises, which persisted despite highly expansionary monetary policies with near-zero policy interest rates. Those papers interpret the crises and subsequent persistent slump as an equilibrium path converging to a liquidity trap steady state. For a policy to exit the liquidity trap, Schmitt-Grohé and Uribe (2017) and Jarocinski and Maćkowiak (2018) propose an exogenous path for the policy interest rate converging to its value in the intended steady state. Under the equilibrium dynamics implied by the liquidity trap, that policy is shown to raise inflation expectations and to stimulate aggregate

[7] Cochrane (2011) criticizes this approach to ruling out equilibria that deviate from the intended steady state (including equilibria involving hyperinflations). Instead, he proposes the specification of policies that are consistent with a unique equilibrium that remains well defined, near or farther away from the intended steady state. A non-Ricardian fiscal policy, combined with a passive monetary policy, is an example of an alternative fiscal–monetary regime that avoids the problems of global multiplicity of the New Keynesian model with an active Taylor rule.

demand and output. The price-level indeterminacy implied by the exogenous path for the nominal rate can be eliminated by a switch to an active fiscal policy. Benigno and Fornaro (2017) develop a model that includes downward nominal wage rigidities and a zero lower bound constraint similar to that of Schmitt-Grohé and Uribe (2017), but in which they embed an endogenous growth mechanism. Under some conditions, two different balanced growth paths may be consistent with equilibrium. One of those paths, which they refer to as a stagnation trap, is characterized by involuntary unemployment and low growth, while the other features high growth and full employment. Expectations about future growth prospects determine which equilibrium obtains.

In this branch of the literature, monetary policy is described by some (generally suboptimal) Taylor-type rule. But the multiplicity of equilibria generated by the zero lower bound is not restricted to that case: as shown in Armenter (2018) and Nakata and Schmidt (2016), it also emerges under the assumption of a central bank optimizing under discretion.[8] Even if such multiplicity is clearly suboptimal, there is little that a central bank operating under discretion can do about it, because the zero lower bound limits its ability to stabilize inflation. As a result, the central bank may find it optimal in some circumstances to accommodate revisions in the private sector's expectations, thus minimizing the damage given the unavoidable deviation from its stabilization targets.

Fiscal Policy and the Zero Lower Bound

In addition to its implications for the design of monetary policy, the zero lower bound also has ramifications for the effects of other shocks, including fiscal policy shocks. This is a consequence of a fairly general principle: In the presence of nominal rigidities, the effects of any fiscal policy intervention are not invariant to the monetary policy rule in place and, more precisely, to the (endogenous) response of nominal and real interest rates to those interventions. By shaping that response, the presence of a zero lower bound constraint has an impact on the effects of fiscal policy shocks.

Eggertsson (2010) and Christiano, Eichenbaum, and Rebelo (2011) analyze the interaction of fiscal policy and the zero lower bound using a New Keynesian model as a reference framework. In particular, Eggertsson (2010) considers an environment in which an adverse demand shock pushes the natural rate into negative territory and makes the zero lower bound binding. In that context, he shows that a reduction in taxes on labor or capital income is expansionary, whereas an increase in government purchases has a strong expansionary effect on output. The reason is that tax cuts (as well as other supply-side policies) generate disinflationary pressures that are not matched by a policy interest rate cut, leading to an increase in the real interest rate and a drop in aggregate demand. On the other hand, an increase

[8] Earlier papers examining optimal discretionary policy under the zero lower bound constraint (for example, Adam and Billi 2007) implicitly make an equilibrium selection by constraining the equilibrium to stay in the neighborhood of the targeted (zero) inflation steady state.

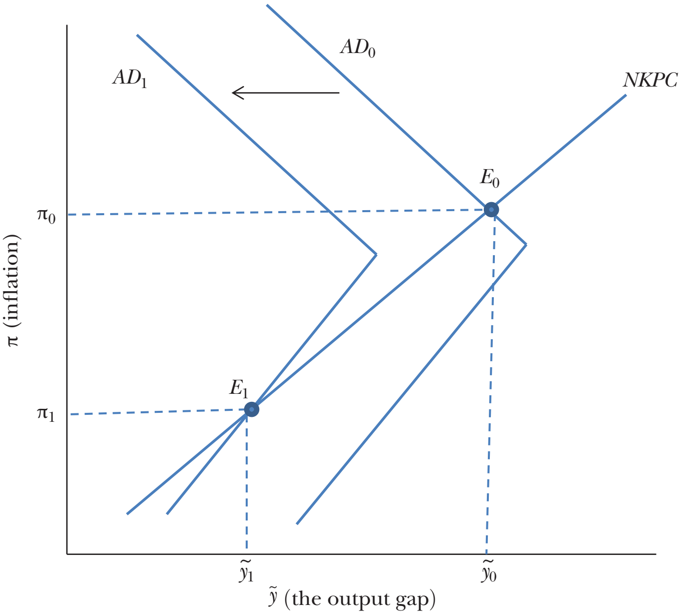
in government purchases has a stronger expansionary effect under a binding zero lower bound than in "normal" times, because the inflationary pressures generated by the fiscal expansion, combined with the absence of a nominal rate adjustment, lead to a drop in the real rate, thus amplifying the effect of the fiscal stimulus.

Christiano, Eichenbaum, and Rebelo (2011) analyze the determinants of the size of the government spending multiplier in connection with a binding zero lower bound. In particular, they show that the multiplier is very sensitive to how long the zero lower bound is expected to be a binding constraint. When they extend the basic New Keynesian model to allow for endogenous capital accumulation, the size of the government spending multiplier becomes even larger, since investment—which is inversely related to the real interest rate—responds procyclically to the fiscal shock thus amplifying the effect of the latter. Using an estimated dynamic stochastic general equilibrium model, they quantify the value of the fiscal multiplier under a binding zero lower bound to be in a neighborhood of 2 for an increase in spending lasting 12 quarters. This contrasts with a multiplier smaller than one when the model is simulated under "normal" times, with the central bank responding to a fiscal expansion according to a conventional Taylor rule.

Figure 3 illustrates the implications of the zero lower bound constraint in the face of a negative aggregate demand shock (for example, resulting from a reduction in government purchases). Note that with the zero lower bound constraint, the AD schedule becomes upward-sloping when inflation reaches a level that makes the zero lower bound constraint binding, given the interest rate rule: Further reductions in inflation raise the real interest rate and lower aggregate demand and the output gap. A leftward shift in the AD schedule, if sufficiently large, pulls the economy into the region in which the decline in inflation cannot be offset by a more-than-proportional reduction in the policy interest rate, thus amplifying the negative impact of the shock on both inflation and the output gap.

The effectiveness of different fiscal policies at stimulating the economy under a binding zero lower bound is not invariant to the reason why the constraint has become binding. In the papers discussed above, the zero lower bound becomes binding as a result of an adverse fundamental shock that is sufficiently large to push the policy rate against the zero lower bound constraint, making it impossible for the central bank to stabilize inflation and the output gap. Mertens and Ravn (2014) focus instead on expectational or nonfundamental liquidity traps, which may emerge as a result of self-fulfilling expectations, due to the global indeterminacy discussed in the previous subsection. In the case of an expectational liquidity trap, they show that an increase in government purchases has a small effect on output (smaller than in normal times), whereas a tax cut is expansionary. The main factor behind those predictions is the differential effect on inflation, which in this model is positive in the case of a tax cut, but negative for a spending increase. Given that these predictions are exactly the opposite to those arising in the case of a fundamental liquidity trap, it follows that a good diagnosis of the nature of a liquidity trap is essential in order to evaluate the effects of a given fiscal policy response.

Figure 3
The Basic New Keynesian Model with Zero Lower Bound Constraint



Note: Figure 3 illustrates the implications of the zero lower bound constraint in the face of a negative aggregate demand shock (for example, resulting from a reduction in government purchases). The AD schedule (after "aggregate demand") combines the dynamic IS equation, the interest rate rule, and the zero lower bound constraint. The NKPC schedule represents the positive relation between the same two variables implied by the New Keynesian Phillips Curve, given inflation expectations. The economy's equilibrium is determined by the intersection of the two schedules.

Heterogeneity

The standard New Keynesian model, like most of its predecessors in the real business cycle literature, represents an economy inhabited by an infinitely-lived representative household. That assumption is obviously unrealistic. But of course, all models involve some simplification of reality so as to focus on the specific issue at hand. In macroeconomics, one specific question is how to explain aggregate fluctuations and their interaction with monetary policy in a relatively compact and tractable manner. It is not immediately obvious why the finiteness of human life is an aspect of reality that will be especially important in building such a model. After all, individuals are heterogenous in their economic behavior along a number of dimensions: education, wealth, income, preference for leisure, risk-taking, perceptions of relevant time horizons, and many more. For tractability, a macroeconomic model will of necessity leave out many of these ways in which people vary. In this spirit, one can perhaps argue that a representative household is a defensible starting point for a macro model.

This section will discuss a growing literature that argues that the representative household assumption is less innocuous than may appear, even when the focus is to understand aggregate fluctuations and macroeconomic policy.[9]

An important problem (though not the only one) that arises with representative household models is that in equilibrium there are neither savers nor borrowers, even in the absence of financial frictions, since everyone is identical. Thus, in order to understand whether the presence and nature of financial frictions have nontrivial implications for economic fluctuations and monetary policy, it is necessary to relax the representative household assumption. A large (and growing) number of papers have undertaken this approach in recent years, using a suitably modified New Keynesian model as a reference framework.

Here, I will focus on the latest generation of New Keynesian models with heterogeneous agents and financial frictions, models generally referred to as HANK (for Heterogeneous Agent New Keynesian).[10] A first feature shared by the recent wave of HANK models, which differentiates them from the baseline New Keynesian model, is the assumption of idiosyncratic shocks to households' labor productivity and hence to their wage. Those shocks are often assumed to follow a stochastic process that is consistent with some features of the microdata. Secondly, it is generally assumed that only a small number of assets can be traded, and that some exogenous borrowing limit exists. As a result, households cannot perfectly insure themselves against idiosyncratic risk. Furthermore, a (time-varying) fraction of households face a binding borrowing constraint, which makes their consumption respond strongly to fluctuations in current income. The previous features imply that no simple dynamic IS equation like the one described earlier can be derived. However, the other two main elements of the basic New Keynesian framework—the New Keynesian Phillips curve and the interest rate rule—are not directly affected by the introduction of heterogeneity.

One of the main lessons emerging from the analysis of heterogenous agent New Keynesian (HANK) models can be summarized as follows: The presence of uninsurable idiosyncratic shocks, combined with the existence of borrowing limits, implies that different households, even if they otherwise appear identical before the shocks arise, may have at any point in time very different marginal propensities to consume. As a result, the macroeconomic effects of any aggregate shock will be amplified or dampened depending on the way the shock (and the changes that it triggers) affects the distribution of income and wealth across households.

Several recent papers provide an insightful analysis of that mechanism. Auclert (2017) studies the different channels through which heterogeneity shapes the effect

[9] The bulk of the recent literature on heterogeneity has focused on the household sector. For an example of the implications of firm-level heterogeneity, see Adam and Weber (2018).

[10] Guerrieri and Lorenzoni (2017), Oh and Reis (2012), and McKay and Reis (2016) were among the first contributions to this literature, focusing, respectively, on the effects of credit crunches, transfers, and automatic stabilizers. Subsequent contributions include Auclert (2017), Kaplan, Moll, and Violante (2018), Ravn and Sterk (2016), Gornemann, Kuester, and Nakajima (2016), Farhi and Werning (2017), Werning (2015), and Debortoli and Galí (2017), among many others.

of an exogenous monetary policy shock on individual and aggregate consumption. Two of those channels are already present in the Representative Agent New Keynesian model (henceforth, RANK, for short): 1) intertemporal substitution, in response to changes in real interest rates; and 2) the change in consumption induced by the resulting changes in aggregate income, which is a source of a multiplier effect. A HANK economy, on the other hand, provides three additional channels that occur as a consequence of the *redistribution* that takes place in response to a monetary policy change: 1) the *earnings heterogeneity* channel is associated with the fact that some households see their income increase more than proportionally to aggregate income, while others lose in relative terms; 2) the *Fisher channel* refers to the fact that different households have at any point in time different net positions in nominal assets, whose real value will be affected by the change in the price level resulting from the monetary policy intervention; and 3) the *unhedged interest rate exposure* channel arises because of likely differences across households in the mismatch between durations of assets and liabilities (including planned consumption among the latter).

A key determinant of the impact of each of the three redistribution channels is given by the size and sign of their covariance with marginal propensities of consumption across households—that is, by the extent to which the redistribution caused by a monetary policy intervention favors households with a relatively high or a relatively low marginal propensity to consume. In principle, those covariances can be estimated using microdata on households' consumption, income, and balance sheets. The evidence reported in Auclert (2017) suggests that such redistribution channels are likely to amplify the effects of monetary policy on aggregate consumption. To see this, consider an expansionary policy that lowers real interest rates and raises output and inflation. We know that households with relatively low income and wealth also tend to have relatively high marginal propensities to consume. Furthermore, those households will tend to benefit more from the expansionary policy for several reasons backed up by micro evidence: 1) their earnings increase more than proportionally during output expansions; 2) they tend to have relatively large negative net nominal asset positions, and hence experience a relatively larger increase in their net wealth (in real terms) when the price level rises; and 3) they have a lower interest rate exposure (because they tend to have high current consumption and debt repayments relative to income) and thus benefit more from the reduction in real interest rates. Thus, by redistributing income and wealth towards households with a high marginal propensity to consume, the three channels above work in the direction of amplifying the response of aggregate consumption to an interest rate reduction. Interestingly, Auclert (2017) also shows that these empirical properties emerge, at least in qualitative terms, as an equilibrium outcome in a standard incomplete markets model (à la Bewley–Hugget–Aiyagari) calibrated to the US economy. Auclert's analysis thus points to the need to introduce realistic heterogeneity in monetary models in order to capture better the effects of monetary policy, though further work is needed to assess empirically the quantitative importance of each of those channels.

Werning (2015) develops a general framework to identify some of the channels through which heterogeneity and incomplete markets imply a departure from the aggregate implications of the standard representative agent approach. For example, he first considers an economy with idiosyncratic risk but no borrowing or lending, and no outside assets (for example, no government debt or physical capital), and in which household income is proportional to aggregate income. In that setting, the relation between aggregate consumption and interest rates turns out to be identical to that in the RANK model. A similar "as if" result holds for an economy with borrowing and lending and outside assets if liquidity is acyclical—that is, if asset prices and/or borrowing limits move in proportion to income.

Werning's (2015) framework can be seen as a useful benchmark to understand the properties of different HANK models in the literature in which some of the above assumptions are relaxed. Two examples, discussed by Werning, illustrate that point. The model in Ravn and Sterk (2016) combines rigidities in price-setting, characteristic of New Keynesian models, with search and matching in the labor market. In that framework, the rise in unemployment resulting from a tightening of monetary policy leads to an increase in precautionary savings and, hence, an amplification of the effects of monetary policy relative to a RANK economy. By contrast, McKay, Nakamura, and Steinsson (2016) analyze a model with idiosyncratic shocks in which the effects of interest changes are dampened relative to the RANK benchmark, as a result of a built-in procyclical earnings risk (caused by the assumption of an even distribution of countercyclical profits among workers) and countercyclical liquidity (resulting from constant government debt). The implied dampening of the response to monetary policy is presented by McKay et al. as a possible explanation for the forward guidance puzzle discussed in the previous section.

A distinct feature of the HANK model developed in Kaplan, Moll, and Violante (2018) is the coexistence of two assets: a low-return liquid asset and a high-return illiquid asset (think of housing) whose conversion into the liquid asset is subject to convex transaction costs. An implication of the latter assumption is the presence, at any point in time, of a sizable fraction of households that are wealthy but consume in a hand-to-mouth fashion, because the bulk of their wealth is held in the illiquid asset. The presence of those households, combined with those who consume hand-to-mouth because they have low incomes (and are subject to borrowing constraints), implies that a large fraction of the population is highly sensitive to labor income shocks (idiosyncratic and aggregate) but not very responsive to interest rate changes. As Kaplan et al. show, this shift dramatically changes the nature of the monetary policy transmission mechanism as compared to the RANK model. In the HANK model, the *direct* effect of changes in the interest rate on consumption (its effect conditional on an unchanged path for aggregate income) is much less important than its *indirect* effect (resulting from the induced changes in aggregate income). That property is in stark contrast with the RANK model, in which the direct effect is overwhelmingly dominant, because the representative household can substitute consumption intertemporally and, under any plausible calibration, will have a very small marginal propensity to consume out of current income.

In addition, the analysis in Kaplan, Moll, and Violante (2018) highlights an important property of HANK models: the aggregate effects of monetary policy shocks (or for that matter, of any other disturbance) will be shaped by the fiscal policy response to it, and, in particular, by the extent and nature of the redistributional effects of that response.

If we accept that some heterogeneity is useful, we still face a question of how much. The Two-Agent New Keynesian model (TANK, for short) is a relatively simple way of introducing heterogeneity. In this approach, a constant fraction of households is assumed to have no access to financial markets and just consume their current labor income, while the remaining fraction can buy and sell assets in an unconstrained way, as in the basic New Keynesian model. There are no other sources of heterogeneity within each type of households. Early applications of the TANK framework include Galí, López-Salido, and Vallés (2007) and Bilbiie (2008). In Debortoli and Galí (2017), my coauthor and I seek to understand the extent to which TANK models can provide a tractable approximation to their HANK counterparts.[11] Both alternatives share a key feature missing from representative agent models, namely, the fact that at any point in time a fraction of agents face a binding borrowing constraint (or behave as if they did), but TANK models assume a constant fraction of constrained agents, rather than allowing that fraction to vary endogenously as in in richer HANK models. Also, TANK models ignore the impact on agents' current decisions of the likelihood of being financially constrained in the future. Finally, credit-constrained households in HANK models have a marginal propensity to consume below one, especially in response to positive shocks, in contrast with hand-to-mouth households in TANK models, whose marginal propensity to consume is one at all times. Of course, the main advantage of TANK models relative to HANK models lies in their tractability, since there is no need to keep track of the wealth distribution and its changes over time.

Perhaps surprisingly, in Debortoli and Galí (2017) we show that a simple TANK model approximates well, both from a qualitative and a quantitative viewpoint, the aggregate dynamics of a canonical HANK model in response to aggregate shocks. Firstly, a properly calibrated TANK model approximates well the heterogeneity of consumption between constrained and unconstrained households. Secondly, for standard calibrations of the HANK model, consumption heterogeneity within the subset of unconstrained households (which the TANK model abstracts from) remains roughly constant, since those agents are able to limit consumption fluctuations by borrowing and saving.

Overlapping Generations

The assumption of an infinitely-lived representative household found in the standard New Keynesian model has implications that go beyond those emphasized

[11] See also Bilbiie (2017) and Bilbiie and Ragot (2017).

in the literature discussed in the previous section. The discussion to this point has focused on the interaction of idiosyncratic shocks and borrowing constraints as a source of heterogeneity in marginal propensities to consume across households, with its consequent implications for the transmission of monetary and fiscal policies. Less discussed but equally important are, in my opinion, other implications of the infinitely-lived representative household assumption.

Firstly, the assumption of an infinitely-lived representative consumer implies a tight link between the real interest rate and the consumer's time discount rate along a balanced growth path. That relation all but rules out the possibility of a persistently negative natural rate of interest, with the consequent challenges that the latter would pose on a price-stability–oriented monetary policy due to the zero lower bound on nominal interest rates. Notice that in the examples from the literature on the zero lower bound discussed earlier, the natural rate is assumed to be negative temporarily and, possibly, recurrently, but not permanently.

Secondly, the assumption of an infinitely-lived representative household rules out the existence of rational bubbles in equilibrium. After all, if a rational bubble exists in equilibrium, it must grow at the rate of interest and must necessarily be in assets held by the representative household. But the optimal path for consumption and savings of a representative household is inconsistent with holding assets in the long-run that grow at the rate of interest, which rules out the possibility of a rational bubble in that environment (for a proof, see Santos and Woodford 1997). On the other hand, there is a widespread view among policymakers and commentators that bubbles, like the housing bubble experienced in the 2000s, can play a role in financial crises and economic fluctuations. There is also a persistent debate about how monetary policy should respond to the emergence of those bubbles. The fact that the New Keynesian model cannot account for the phenomenon of bubbles seems like a potentially important shortcoming of that framework.

Several recent papers have sought to overcome the limitations of the infinitely-lived household assumption by introducing overlapping generations of finitely-lived individuals in models with nominal rigidities. Thus, Eggertsson et al. (2017) develop a "quantitative" overlapping generations framework with nominal rigidities in order to understand the sources of the decline in the natural rate of interest in the US economy, and to analyze the implications of that decline for monetary policy.[12]

In Galí (2014, 2017a), I develop two alternative models with overlapping generations, monopolistic competition, and sticky prices, and show that asset price bubbles may emerge in equilibria.[13] In both models, a necessary condition for the existence of such bubbly equilibria is a natural rate of interest below the balanced growth path for

[12] A permanent negative natural rate may also arise in models with infinitely-lived agents in the presence of heterogeneity and incomplete markets. For example, Auclert and Rognlie (2018) provide a model to illustrate that possibility.

[13] In Galí (2014), I assume a two-period-lived households and an inelastic labor supply (implying a constant output in equilibrium). Bubble fluctuations imply a stochastic redistribution of consumption across cohorts. By contrast, in Galí (2017a), I introduce asset price bubbles in a perpetual youth model à la Blanchard–Yaari, in which individuals die with a constant probability.

the economy. When bubbles exist, changes in their size can generate aggregate fluctuations, even in the absence of shocks to fundamentals. In that context, one can analyze the implications of alternative monetary policy rules on fluctuations and welfare, since the evolution of bubbles is not independent of the interest rate. A central message of both papers is that a "leaning against the bubble" monetary policy, modeled as an interest rate rule that includes the size of the bubble as one of its arguments, is generally suboptimal and dominated by a policy that focuses on stabilizing inflation.[14]

The Road Ahead

The standard New Keynesian framework as it existed a decade ago has faced challenges in the aftermath of the financial crisis of 2007–2009. Much of the work extending that framework over the past few years has aimed at overcoming some of those challenges. In the present paper, I have described a sample of recent research that extends the standard New Keynesian framework along different dimensions, with a focus on adapting it to take into account the zero lower bound constraint on the nominal interest rates, and household heterogeneity. [15]

However, none of the extensions of the New Keynesian model proposed in recent years seem to capture an important aspect of most financial crises—namely, a gradual build-up of financial imbalances leading to an eventual "crash" characterized by defaults, sudden-stops of credit flows, asset price declines, and a large contraction in aggregate demand, output, and employment. Most of the extensions found in the literature share with their predecessors a focus on equilibria that take the form of stationary fluctuations driven by exogenous shocks. This is also the case in variants of those models that allow for financial frictions of different kinds (for example, Bernanke, Gertler, and Gilchrist 1999; Christiano, Motto, and Rostagno 2014). In those models, financial frictions often lead to an amplification of the effects of nonfinancial shocks. Also, the presence of financial frictions can lead to additional sources of fluctuations: for example, via risk shocks in Christiano et al. (2014) or exogenous changes in the tightness of borrowing constraints in Guerrieri and Lorenzoni (2017). Overall, it's fair to say that most attempts to use a version of the New Keynesian models to explain the "financial crisis" end up relying on a *large exogenous shock* that impinges on the economy unexpectedly, triggering a large recession, possibly amplified by a financial accelerator mechanism embedded in the model.

There have been a few attempts to model economies that are less subject to the previous criticism. As one example, Boissay, Collard, and Smets (2016) analyze a real model with asymmetric information in the interbank market, in which a sequence of small shocks may pull an economy towards a region with multiple equilibria,

[14] Nonrational bubbles may exist also in economies with an infinitely-lived representative household. For example, see Adam and Woodford (2013) for an analysis of optimal policy in the context of a New Keynesian model with nonrational housing bubbles.
[15] The discussion in this section draws heavily on Galí (2017b).

including equilibria characterized by a freeze in the interbank market, a credit crunch, and a prolonged recession. A monetary extension of such a framework would seem highly welcome.

As another example, in Galí (2017a), I explore the possibility of fluctuations driven by stochastic bubbles in a New Keynesian model with overlapping generations. Stochastic bubbles grow at a rate above the long-term growth of the economy, generating a boom in output and employment. But these bubbles may collapse at any time with some (exogenously given) probability, pulling down aggregate demand and output when they do. Despite the highly stylized nature of the model, the implied equilibrium appears consistent with the pattern of asset price booms followed by sudden busts (and the induced recession) that has characterized historical financial crises. However, the framework abstracts from financial frictions and, in particular, from the important role that high credit growth seems to have played in bubble episodes (Jordà, Schularick, and Taylor 2015). It also leaves unexplained the factors that ultimately drive the innovations in the aggregate bubble, as well as its eventual bursting.

As yet another example, Basu and Bundick (2017) analyze a nonlinear version of the New Keynesian model where large and persistent slumps may arise as a result of the strong feedback between aggregate demand and (endogenous) volatility, resulting from the interaction of precautionary savings and a zero lower bound constraint. With a zero lower bound constraint, there is no guarantee that the central bank will manage to stabilize the economy *on the downside*, which in turn raises households' perceived volatility of future consumption (as well as its negative skewness), leading to higher precautionary savings, a reduction in output, a higher probability of falling into a liquidity trap and an additional feedback effect on volatility and skewness.

These are only examples of efforts to introduce mechanisms that may generate patterns that one may relate, at least qualitatively, to those observed in actual financial crises. In the years ahead, I expect further research along these lines, incorporating stronger endogenous propagation mechanisms that may help account for large and persistent fluctuations in output and inflation without the need to rely on large (and largely unexplained) exogenous shocks.

But in the meantime, New Keynesian economics is alive and well. The New Keynesian model has proved to be quite flexible, with a growing number of extensions being developed by researchers in order to incorporate new assumptions or account for new phenomena. Indeed, it is hard to think of an alternative macroeconomic paradigm that would do away with the two defining features of the New Keynesian model: nominal rigidities and monetary non-neutralities.

■ *I am grateful to Klaus Adam, Adrien Auclert, Davide Debortoli, Mark Gertler, Gordon Hanson, Greg Kaplan, and Stephanie Schmitt-Grohé for comments, and to Timothy Taylor for a top-notch editorial job. Christian Höynck and Cristina Manea provided excellent research assistance. I have benefited from the financial support of the Generalitat de Catalunya (AGAUR Grant 2017SGR1393) and the Ministerio de Economía, Industria y Competitividad (Grant ECO2017-87827).*

References

Adam, Klaus, and Roberto Billi. 2006. "Optimal Monetary Policy under Commitment with a Zero Bound on Nominal Interest Rates." *Journal of Money, Credit and Banking* 38(7): 1877–1905.

Adam, Klaus, and Roberto Billi. 2007. "Discretionary Monetary Policy and the Zero Lower Bound on Nominal Interest Rates." *Journal of Monetary Economics* 54(3): 728–52.

Adam, Klaus, and Henning Weber. 2018. "Optimal Trend Inflation." February, http://adam.vwl.uni-mannheim.de/fileadmin/user_upload/adam/Adam_Weber_Optimal_Inflation.pdf.

Adam, Klaus, and Michael Woodford. 2013. "Housing Prices and Robustly Optimal Monetary Policy." June 29, http://adam.vwl.uni-mannheim.de/fileadmin/user_upload/adam/research/AW_AssetPrices.pdf.

Angeletos, George-Marios, and Chen Lian. Forthcoming. "Forward Guidance without Common Knowledge." *American Economic Review.*

Armenter, Roc. 2018. "The Perils of Nominal Targets." *Review of Economic Studies* 85(1): 50–86.

Aruoba, Boragan, Pablo Cuba-Borda, and Frank Schorfheide. 2018. *Review of Economic Studies* 85(1): 87–118.

Auclert, Adrien. 2017. "Monetary Policy and the Redistribution Channel." NBER Working Paper 23451.

Auclert, Adrien, and Matthew Rognlie. 2018. "Inequality and Aggregate Demand." NBER Working Paper 24280.

Barro, Robert J., and David B. Gordon. 1983. "A Positive Theory of Monetary Policy in a Natural Rate Model." *Journal of Political Economy* 91(4): 589–610.

Basu, Susanto, and Brent Bundick. 2017. "Uncertainty Shocks in a Model of Effective Demand." *Econometrica* 85(3): 937–58.

Basu, Susanto, John G. Fernald, and Miles S. Kimball. 2006. "Are Technology Improvements Contractionary?" *American Economic Review* 96(5): 1418–48.

Benhabib, Jess, Stephanie Schmitt-Grohé, and Martín Uribe. 2001. "The Perils of Taylor Rules." *Journal of Economic Theory* 96(1–2): 40–69.

Benhabib, Jess, Stephanie Schmitt-Grohé, and Martín Uribe. 2002. "Avoiding Liquidity Traps." *Journal of Political Economy* 110(3): 535–63.

Benigno, Gianluca, and Luca Fornaro. 2017. "Stagnation Traps." *Review of Economic Studies*, forthcoming.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." Chap. 21 in *Handbook of Macroeconomics,* vol. 1C, edited John B. Taylor and Michael Woodford. Amsterdam: Elsevier, pp. 1341–93.

Bilbiie, Florin O. 2008. "Limited Asset Market Participation, Monetary Policy and (Inverted) Aggregate Demand Logic." *Journal of Economic Theory* 140(1): 162–96.

Bilbiie, Florin O. 2017. "The New Keynesian Cross: Understanding Monetary Policy with Hand-to-Mouth Households." CEPR Discussion Paper 11989.

Bilbiie, Florin O., and Xavier Ragot. 2017. "Optimal Monetary Policy and Liquidity with Heterogeneous Households." CEPR Discussion Paper DP11814.

Blanchard, Olivier J., Christopher J. Erceg, and Jesper Lindé. 2016. "Jump-Starting the Euro Area Recovery: Would a Rise in Core Fiscal Spending Help the Periphery?" *NBER Macroeconomics Annual,* vol. 31, pp. 103–82.

Blanchard, Olivier, and Jordi Galí. 2007. "Real Wage Rigidities and the New Keynesian Model." *Journal of Money, Credit and Banking* 39(s1): 35–66.

Blanchard, Olivier, and Jordi Galí. 2010. "Labor Markets and Monetary Policy: A New Keynesian Model with Unemployment." *American Economic Journal: Macroeconomics* 2(2): 1–30.

Boissay, Frédéric, Fabrice Collard, and Frank Smets. 2016. "Booms and Banking Crises." *Journal of Political Economy* 124(2): 489–538.

Bullard, James, and Kaushik Mitra. 2002. "Learning about Monetary Policy Rules." *Journal of Monetary Economics* 49(6): 1105–30.

Calvo, Guillermo A. 1983. "Staggered Prices in a Utility Maximizing Framework." *Journal of Monetary Economics* 12(3): 383–98.

Carlstrom, Charles T., Timothy S. Fuerst, and Matthias Paustian. 2015. "Inflation and Output in New Keynesian Models with a Transient Interest Rate Peg." *Journal of Monetary Economics* 76: 230–43.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 1999. "Monetary Policy Shocks: What Have We Learned and to What End?" In *Handbook of Macroeconomics* vol. 1A, edited by John B. Taylor and Michael Woodford, 65–148.

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

Christiano, Lawrence J., Roberto Motto, and Massimo Rostagno. 2014. "Risk Shocks." *American Economic Review* 104(1): 27–65.

Clarida, Richard, Jordi Galí, and Mark Gertler. 1999. "The Science of Monetary Policy: A New

Keynesian Perspective." *Journal of Economic Literature* 37(2): 1661–1707.

Clarida, Richard, Jordi Galí, and Mark Gertler. 2000. "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory." *Quarterly Journal of Economics* 115(1): 147–80.

Clarida, Richard, Jordi Galí, and Mark Gertler. 2002. "A Simple Framework for International Monetary Policy Analysis." *Journal of Monetary Economics* 49(5): 879–904.

Cochrane, John H. 2011. "Determinacy and Identification with Taylor Rules." *Journal of Political Economy* 119(3): 565–615.

Debortoli, Davide, and Jordi Galí. 2017. "Monetary Policy with Heterogeneous Agents: Insights from TANK Models." CREI Working Paper, http://www.crei.cat/wp-content/uploads/2018/03/dg_tank.pdf.

Debortoli, Davide, and Aeimit Lakdawala. 2016. "How Credible is the Federal Reserve? A Structural Estimation of Policy Re-Optimizations." *American Economic Journal: Macroeconomics* 8(3): 42–76.

Del Negro, Marco, Gauti Eggertsson, Andrea Ferrero, and Nobuhiro Kiyotaki. 2017. "The Great Escape? A Quantitative Evaluation of the Fed's Liquidity Facilities." *American Economic Review* 107(3): 824–57.

Del Negro, Marco, Marc Giannoni, and Christina Patterson. 2012. "The Forward Guidance Puzzle." Staff Report 574, Federal Reserve Bank of New York.

Del Negro, Marco, and Frank Schorfheide. 2013. "DSGE Model-Based Forecasting." In *Handbook of Economic Forecasting*, vol. 2A, edited by Graham Elliot and Allan Timmermman, 58–137. Elsevier B.V.

Eggertsson, Gauti. 2010. "What Fiscal Policy is Effective at Zero Interest Rates?" *NBER Macroeconomics Annual*, vol. 25, pp. 59–112.

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

Eggertson, Gauti B. and Michael Woodford. 2003. "The Zero Bound on Interest Rates and Optimal Monetary Policy." *Brookings Papers on Economic Activity*, vol. 1, pp. 139–211.

Erceg, Christopher J., Dale W. Henderson, and Andrew T. Levin. 2000. "Optimal Monetary Policy with Staggered Wage and Price Contracts." *Journal of Monetary Economics* 46(2): 281–313.

Evans, George W., and Seppo Honkapohja. 2003. "Expectations and the Stability Property for Optimal Monetary Policies." *Review of Economic Studies* 70(4): 807–24.

Farhi, Emmanuel, and Iván Werning. 2017. "Monetary Policy, Bounded Rationality and Incomplete Markets." NBER Working Paper 23281.

Gabaix, Xavier. 2017. "A Behavioral New Keynesian Model." NBER Working Paper 22954.

Galí, Jordi. 1999. "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?" *American Economic Review* 89(1): 249–71.

Galí, Jordi. 2014. "Monetary Policy and Rational Asset Price Bubbles." *American Economic Review* 104(3): 721–52.

Galí, Jordi. 2015. *Monetary Policy, Inflation and the Business Cycle: An Introduction to the New Keynesian Framework and Its Applications*, 2nd edition. Princeton University Press.

Galí, Jordi. 2017a. "Monetary Policy and Bubbles in a New Keynesian Model with Overlapping Generations." CREI Working Paper. http://www.crei.cat/wp-content/uploads/2018/01/nk_bubbles_dec2017.pdf

Galí, Jordi. 2017b. "Some Scattered Thoughts on DSGE Models." In *DSGE Models in the Conduct of Policy; Use as Intended*, edited by Refet S. Gürkaynak and Cédric Tille. CEPR Press, 86–92.

Galí, Jordi, J. David López-Salido, and Javier Vallés. 2007. "Understanding the Effects of Government Spending on Consumption." *Journal of the European Economic Association* 5(1): 227–70.

Galí, Jordi, and Tommaso Monacelli. 2005. "Monetary Policy and Exchange Rate Volatility in a Small Open Economy." *Review of Economic Studies* 72(3): 707–34.

Gornemann, Nils, Keith Kuester, and Makoto Nakajima. 2016. "Doves for the Rich, Hawks for the Poor? Distributional Consequences of Monetary Policy." International Finance Discussion Papers 1167, Board of Governors of the Federal Reserve System.

Guerrieri, Veronica, and Guido Lorenzoni. 2017. "Credit Crises, Precautionary Savings, and the Liquidity Trap." *Quarterly Journal of Economics* 132(3): 1427–67.

Jarociński, Marek, and Bartosz Maćkowiak. 2018. "Monetary Fiscal Interactions and the Euro Area's Malaise." *Journal of International Economics* 112: 251–66.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2015. "Leveraged Bubbles." *Journal of Monetary Economics* 76(Supplement): S1–S20.

Jung, Taehun, Yuki Teranishi, and Tsutomo Watanabe. 2005. "Optimal Monetary Policy at the Zero-Interest-Rate Bound." *Journal of Money, Credit and Banking* 37(5): 813–35.

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante. 2018. "Monetary Policy according to HANK." *American Economic Review* 108(3): 697–743.

Keynes, John Maynard. 1936. *The General Theory* *of Employment, Interest, and Money.* Cambridge University Press.

Kydland, Finn E., and Edward C. Prescott. 1980. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy* 85(3): 473–92.

Lindé, Jesper, Frank Smets, and Raf Wouters. 2016. "Challenges for Central Banks' Macro Models." Chap. 29 in *Handbook of Macroeconomics* vol. 2B, edited by John B. Taylor and Harald Uhlig, 2186–2262. Elsevier B.V.

McKay, Alisdair, Emi Nakamura, and Jón Steinsson. 2016. "The Power of Forward Guidance Revisited." *American Economic Review* 106(10): 3133–58.

McKay, Alisdair, Emi Nakamura, and Jón Steinsson. 2017. "The Discounted Euler Equation: A Note." *Economica* 84(336): 820–31.

McKay, Alisdair, and Ricardo Reis. 2016. "The Role of Automatic Stabilizers in the U.S. Business Cycle." *Econometrica* 84(1): 141–194.

Mertens, Karel R. S., and Morten O. Ravn. 2014. "Fiscal Policy in an Expectations-Driven Liquidity Trap." *Review of Economic Studies* 81(4): 1637–67.

Nakata, Taisuke, and Sebastian Schmidt. 2016. "Gradualism and Liquidity Traps." Finance and Economics Discussion Series 2016-092, Board of Governors of the Federal Reserve System (US).

Nakov, Anton. 2008. "Optimal and Simple Monetary Policy Rules with Zero Floor on the Nominal Interest Rate." *International Journal of Central Banking* 4(2): 73–127.

Oh, Hyunseung, and Ricardo Reis. 2012. "Targeted Transfers and the Fiscal Response to the Great Recession." *Journal of Monetary Economics* 59(Supplement): S50–S64.

Ravn, Morten, and Vincent Sterk. 2016. "Macroeconomic Fluctuations with HANK & SAM: An Analytical Approach." CEPR Discussion Paper 11696.

Rogoff, Kenneth. 1985. "The Optimal Degree of Commitment to an Intermediate Monetary Target." *Quarterly Journal of Economics* 100(4): 1169–89.

Santos, Manuel S., and Michael Woodford. 1997. "Rational Asset Pricing Bubbles." *Econometrica* 65(1): 19–57.

Schamburg, Ernst, and Andrea Tambalotti. 2007. "An Investigation of the Gains from Commitment in Monetary Policy." *Journal of Monetary Economics* 54(2): 302–324.

Schmitt-Grohé, Stephanie, and Martín Uribe. 2017. "Liquidity Traps and Jobless Recoveries." *American Economic Journal: Macroeconomics* 9(1): 165–204.

Smets, Frank, Kai Christoffel, Gunter Coenen, Roberto Motto, and Massimo Rostagno. 2010. "DSGE Models and Their Use at the ECB." *SERIEs: Journal of the Spanish Economic Associsation* 1(1–2): 51–65.

Steinsson, Jon. 2003. "Optimal Monetary Policy in an Economy with Inflation Persistence." *Journal of Monetary Economics* 50(7): 1425–56.

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

Trigari, Antonella. 2009. "Equilibrium Unemployment, Job Flows, and Inflation Dynamics." *Journal of Money, Credit and Banking* 41(1): 1–33.

Vestin, David. 2006. "Price-level versus Inflation Targeting." *Journal of Monetary Economics* 53(7): 1361–76.

Werning, Iván. 2015. "Incomplete Markets and Aggregate Demand." NBER Working Paper 21448.

Woodford, Michael. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy.* Princeton University Press.

Woodford, Michael. 2010. "Robustly Optimal Monetary Policy with Near-Rational Expectations." *American Economic Review* 100(1): 274–303.

Woodford, Michael. 2013. "Forward Guidance by Inflation-Targeting Central Banks." http://www.columbia.edu/~mw2230/RiksbankIT.pdf.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 113–140

On DSGE Models

Lawrence J. Christiano, Martin S. Eichenbaum, and Mathias Trabandt

The outcome of any important macroeconomic policy change is the net effect of forces operating on different parts of the economy. A central challenge facing policymakers is how to assess the relative strength of those forces. Economists have a range of tools that can be used to make such assessments. Dynamic stochastic general equilibrium (DSGE) models are the leading tool for making such assessments in an open and transparent manner.

To be concrete, suppose we are interested in understanding the effects of a systematic change in policy, like switching from inflation targeting to price-level targeting. The most compelling strategy would be to do randomized control trials on actual economies, but that course of action is not available to us. So what are the alternatives? It is certainly useful to study historical episodes in which such a similar policy switch occurred or to use reduced-form time series methods, but these approaches also have obvious limitations. In the historical approach, the fact that no two episodes are exactly the same always raises questions about the relevance of a past episode for the current situation. In the case of reduced-form methods, it is not always clear which parameters should be changed and which should be kept constant across policy options. Inevitably, assessing the effects of a systematic policy change has to involve the use of a model.

■ *Lawrence J. Christiano is the Alfred W. Chase Professor of Economics and Martin S. Eichenbaum is the Charles Moskos Professor of Economics, both at Northwestern University, Evanston, Illinois. Mathias Trabandt is Professor of Macroeconomics at the School of Business and Economics, Freie Universität Berlin, Berlin, Germany and Research Fellow at the Halle Institute for Economic Research, Halle, Germany. Their email addresses are l-christiano@northwestern.edu, eich@northwestern.edu, and mathias.trabandt@gmail.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.113 doi=10.1257/jep.32.3.113

To be useful for policy analysis, dynamic stochastic general equilibrium models must be data-based. As a practical matter, macroeconomic data are not sufficient for discriminating between many alternative models that offer different answers to policy questions. Put differently, many DSGE models are observationally equivalent with respect to macro data. But modern DSGE models are based on microeconomic foundations. So microeconomic data and institutional facts can be brought to bear on their design, construction, and evaluation. Micro data break the observational equivalence that was the bane of macroeconomists.

The openness and transparency of dynamic stochastic general equilibrium models is a virtue—but it also makes them easy to criticize. Suspicious assumptions can be highlighted. Inconsistencies with the evidence can easily be spotted. Forces that are missing from the model can be identified. The process of responding to informed criticisms is a critical part of the process of building better DSGE models. Indeed, the transparent nature of DSGE models is exactly what makes it possible for diverse groups of researchers—including those who don't work on DSGE models—to be part of the DSGE project.

Some analysts object to working with dynamic stochastic general equilibrium models and prefer instead to think about policy by working with small equilibrium models that emphasize different subsets of the economy, labor, or financial markets. This approach has a vital contribution to make, because small models help build intuition about the mechanisms at work in DSGE models. But this approach cannot be a substitute for DSGE models themselves, because quantitative conclusions about the overall economic impact of a policy requires informal judgment as one integrates across individual small-scale models. The small-model approach to policy thus involves implicit assumptions and lacks the transparency of the DSGE approach.

To be clear, policy decisions are made by real people using their best judgment. Used wisely, dynamic stochastic general equilibrium models can improve and sharpen that judgment. In an ideal world, we will have both wise policymakers and empirically plausible models. But to rephrase Fischer's (2017) quoting of Samuelson on Solow: "We'd rather have Stanley Fischer than a DSGE model, but we'd rather have Stanley Fischer with a DSGE model than without one."

In the next section, we review the state of mainstream dynamic stochastic general equilibrium models before the financial crisis and the Great Recession. We then describe how DSGE models are estimated and evaluated. We address the question of why DSGE modelers—like most other economists and policymakers—failed to predict the financial crisis and the Great Recession, and how DSGE modelers responded to the financial crisis and its aftermath. We discuss how current DSGE models are actually used by policymakers. We then provide a brief response to some criticisms of DSGE models, with special emphasis on Stiglitz (2017), and offer some concluding remarks.

Before the Storm

In this section, we describe early dynamic stochastic general equilibrium models and how they evolved prior to the crisis.

Early Dynamic Stochastic General Equilibrium Models

As a practical matter, people often use the term "dynamic stochastic general equilibrium model" to refer to quantitative models of growth or business cycle fluctuations. A classic example of a quantitative DSGE model is the real business cycle model associated with Kydland and Prescott (1982) and Long and Plosser (1983). These early real business cycle models imagined an economy populated by households who participate in perfectly competitive goods, factor, and asset markets. These models took the position that fluctuations in aggregate economic activity are an efficient response of the economy to the one source of uncertainty in agents' environment, exogenous technology shocks. The associated policy implications are clear: there is no need for any form of government intervention. In fact, government policies aimed at stabilizing the business cycle are welfare reducing.

Excitement about real business cycle models crumbled under the impact of three forces. First, micro data cast doubt on some of the key assumptions of the model. These assumptions include, for example, perfect credit and insurance markets, as well as perfectly frictionless labor markets in which fluctuations in hours worked reflect movements along a given labor supply curve or optimal movements of agents in and out of the labor force (Chetty, Guren, Manoli, and Weber 2011).

Second, the models had difficulty in accounting for some key properties of the aggregate data, such as the observed volatility in hours worked, the equity premium, the low co-movement of real wages and hours worked (Christiano and Eichenbaum 1992; King and Rebelo 1999). Open-economy versions of these models also failed to account for key observations such as the cyclical co-movement of consumption and output across countries (Backus, Kehoe, and Kydland 1992) and the extremely high correlation between nominal and real exchange rates (Mussa 1986).

Third, because money plays no role in real business cycle models, those models seem inconsistent with mainstream interpretations of various historical episodes. An example is Hume's (1742) description of how money from the New World affected the European economy. A different example is the view that the earlier a country abandoned the gold standard during the Great Depression, the sooner its recovery began (Bernanke 1995). A final example is the view that the severity of the US recession in the early 1980s was in large part caused by monetary policy.

Finally, the simple real business cycle model is effectively mute on a host of policy-related questions of vital importance to macroeconomists and policymakers. Examples include: what are the consequences of different monetary policy rules for aggregate economic activity, what are the effects of alternative exchange rate regimes, and what regulations should we impose on the financial sector?

New Keynesian Models

Prototypical pre-crisis dynamic stochastic general equilibrium models built upon the chassis of the real business cycle model to allow for nominal frictions, both in labor and goods markets. These models are often described as New Keynesian DSGE models, but it would be just as appropriate to refer to them as Friedmanite DSGE models. The reason is that they embody the fundamental worldview articulated in Friedman's (1968) seminal Presidential Address to the American Economic

Association. According to this view, hyperinflations aside, monetary policy has essentially no impact on real variables like output and the real interest rate in the long run. However, due to sticky prices and wages, monetary policy matters in the short run.[1] Specifically, a policy-induced transitory fall in the nominal interest rate is associated with a decline in the real interest rate, an expansion in economic activity, and a moderate rise in inflation.

Models in which *permanent* changes in monetary policy induce roughly one-to-one changes in inflation and the nominal rate of interest are said to satisfy the Fisherian property. Models in which *transitory* changes in monetary policy induce movements in nominal interest rates and inflation of the opposite sign are said to satisfy the anti-Fisherian property. The canonical New Keynesian models of Yun (1996), Clarida, Galí, and Gertler (1999), and Woodford (2003) satisfy both properties.

The basic intuition behind the anti-Fisherian property of the New Keynesian model is as follows. Firms set their prices on the basis of current and future marginal costs. The future state of the economy is relatively unaffected by a transitory monetary policy shock, so actual inflation responds relatively little to a policy-induced transitory fall in the nominal interest rate. As a result, the real interest rate declines. Intertemporal substitution by households then induces a rise in current consumption, leading to a rise in labor income. That increase reinforces the contemporaneous rise in consumption and employment. The expansion in employment drives wages and marginal costs up. The latter effect drives inflation up. Because inflation and the nominal interest rate move in opposite directions, the model has the anti-Fisherian property. Less surprisingly, standard New Keynesian models satisfy the Fisherian property because their long-run properties are roughly the same as the underlying real business cycle chassis.

Many researchers found New Keynesian models attractive because they seemed sensible and they allowed researchers to engage in the types of policy debates about which real business cycle models had been silent. A critical question was: What properties should quantitative versions of these models have? To address this question, the empirical literature focused on quantifying the dynamic effects of a shock to monetary policy. This type of shock has long been of interest to macroeconomists. For example, Friedman and Schwartz (1963) attributed the major portion of business cycle variations to exogenous shocks in the money supply. The recent literature finds these shocks interesting because they provide a potentially powerful diagnostic for discriminating between models. Perhaps the most extreme example is that a real business cycle model implies nothing happens to real variables after a monetary policy shock. In contrast, simple New Keynesian models imply that real variables do respond to a monetary policy shock.

A monetary policy shock can reflect a variety of factors, including measurement error in the real-time data on which policymakers condition their actions and

[1] For example, Friedman (1968, p. 10) writes that after the monetary authority increases money growth, "much or most of the rise in income will take the form of an increase in output and employment rather than in prices. People have been expecting prices to be stable, and prices and wages have been set for some time in the future on that basis. It takes time for people to adjust to a new state of demand. Producers will tend to react to the initial expansion in aggregate demand by increasing output, employees by working longer hours, and the unemployed, by taking jobs now offered at former nominal wages."

the basic randomness that is inherent in group decisions. In a seminal paper, Sims (1986) argued that one should identify monetary policy shocks with disturbances to a monetary policy reaction function in which the policy instrument is a short-term interest rate. Bernanke and Blinder (1992) and Christiano, Eichenbaum, and Evans (1996, 1999) identify monetary policy shocks using the assumption that they have no contemporaneous impact on inflation and output.[2] This set of identifying restrictions, like the entire New Keynesian enterprise, falls squarely in the Friedman worldview. In testimony before Congress, Friedman (1959) said: "Monetary and fiscal policy is rather like a water tap that you turn on now and that then only starts to run 6, 9, 12, 16 months from now."

In practice, this Friedman-style identifying strategy is implemented using a vector autoregression representation with a large set of variables. Figure 1, taken from Christiano, Trabandt, and Walentin (2010), displays the effects of identified monetary policy shocks estimated using data covering the period 1951:Q1 to 2008:Q4. For convenience, we only show the response functions for a subset of the variables in the vector autoregression. The dashed lines correspond to 95 percent confidence intervals about the point estimates, shown by the thick solid line.

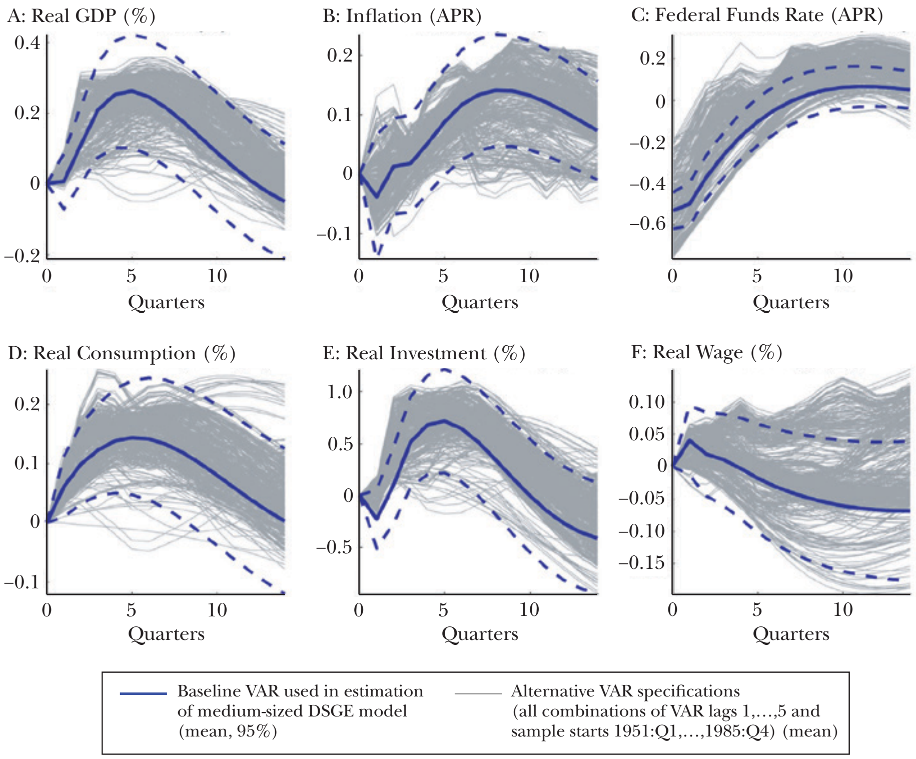
Overall, the results are consistent with the view that an expansionary monetary policy shock has the effects that Friedman (1968) asserted in his Presidential Address. Specifically, an expansionary monetary policy shock corresponding to a decline in the US federal funds rate leads to hump-shaped expansions in consumption, investment, and output, as well as relatively small rises in real wages and inflation. Since the inflation rate moves very little in response to a monetary policy shock, the responses in the real interest rate and the federal funds rate are roughly the same.

A natural question is how robust the results in Figure 1 are to the various technical assumptions underlying the statistical analysis. Here, we focus on sensitivity to the number of lags in the vector autoregression and to the start of the sample period. A vector autoregression represents each variable as a function of the lagged values of all the variables in the system. Denote the number of lags by n. The baseline specification in Figure 1 assumes $n = 2$. Figure 1 reports the results of redoing the analysis for $n = 1, \dots, 5$. For each value of n, Figure 1 reports the results based on starting the sample period in each of the quarters from 1951:Q1 up through 1985:Q4. In this way, we generate 700 sets of results, each of which is displayed by a thin grey line in Figure 1. Note that the basic qualitative properties of the benchmark analysis are remarkably robust, although there are of course specifications of n and the sample period that yield different implications. It is interesting how similar the shape of the confidence and sensitivity intervals are.

In recent years, researchers have developed alternative procedures for identifying monetary policy shocks. These procedures focus on movements in the federal funds futures rate in a tight window of time around announcements made by monetary policymakers: for example, see Gertler and Karadi (2015) who build

[2] Christiano, Eichenbaum, and Evans (1999) show that the results from imposing this assumption on monthly or quarterly data are qualitatively similar. The assumption is obviously more compelling for monthly data.

Figure 1
Vector Autoregression (VAR) Impulse Responses to a Monetary Policy Shock



Source: Christiano, Trabandt, and Walentin (2010).
Note: The figure displays the effects of identified monetary policy shocks estimated using data covering the period 1951:Q1 to 2008:Q4. All data are expressed in deviations from what would have happened in the absence of the shock. The units are given in the titles of the subplots. Percent means percent deviation from unshocked path. APR means annualized percentage rate deviation from the unshocked path. The dashed lines correspond to 95 percent confidence intervals about the point estimates, shown with a thick solid line. The baseline specification of the vector autoregression assumes the number of lags $n = 2$. The figure also reports the results of redoing the analysis for $n = 1, \dots, 5$. For each value of n, the figure reports the results based on starting the sample period in each of the quarters from 1951:Q1 up through 1985:Q4. In this way, we generate 700 sets of results, each of which is displayed by a thin grey line (for details, see Christiano, Trabandt, and Walentin 2010).

on the work of Kuttner (2001) and Gürkaynak, Sack, and Swanson (2005). Broadly speaking, this literature reaches the same conclusions about the effects of monetary policy shocks displayed in Figure 1. In our view, these conclusions summarize the conventional view about the effects of a monetary policy shock.

The Christiano, Eichenbaum, and Evans Model

A key challenge was to develop an empirically plausible version of the New Keynesian model that could account quantitatively for the type of impulse response functions displayed in Figure 1. Christiano, Eichenbaum, and Evans (2005) developed a version of the New Keynesian model that met this challenge. We go into some detail describing the basic features of that model because they form the core of leading pre-crisis dynamic stochastic general equilibrium models, such as Smets and Wouters (2003, 2007).

Consumption and Investment Decisions by the Representative Household. Consistent with a long tradition in macroeconomics, the model economy in Christiano, Eichenbaum, and Evans (2005) is populated by a representative household. At each date, the household allocates money to purchases of financial assets, as well as consumption and investment goods. The household receives income from wages, from renting capital to firms, and from financial assets, all net of taxes.

As in the simple New Keynesian model, Christiano, Eichenbaum, and Evans (2005) make assumptions that imply the household's borrowing constraints are not binding, so the interest rate determines the intertemporal time pattern of consumption. Of course, the present value of income determines the level of consumption. Holding interest rates constant, the solution to the household problem is consistent with a key prediction of Friedman's permanent income hypothesis: persistent changes in income have a much bigger impact on household consumption than transitory changes.

To be consistent with the response of consumption and the interest rate to a monetary policy shock observed in Figure 1, Christiano, Eichenbaum, and Evans (2005) depart from the standard assumption that utility is time-separable in consumption. Generally speaking, that assumption implies that after a policy-induced decline in the interest rate, consumption jumps immediately and then falls. But this pattern is very different from the hump-shape response that we see in Figure 1. To remedy this problem, Christiano, Eichenbaum, and Evans (2005) follow Fuhrer (2000) by adopting the assumption of habit-formation in consumption. Under this specification, the marginal utility of current consumption depends positively on the level of the household's past consumption. Households then choose to raise consumption slowly over time, generating a hump-shape response-pattern as in Figure 1. As it turns out, there is substantial support for habit persistence in the finance, growth, and psychology literatures.[3]

To be consistent with the hump-shaped response of investment to a monetary policy shock, Christiano, Eichenbaum, and Evans (2005) had to assume that households face costs of changing the rate of investment. To see why, note that absent uncertainty, arbitrage implies that the one-period return on capital is equal to the real rate of interest on bonds. Absent any adjustment costs, the one-period return on capital is the sum of the marginal product of capital plus one minus the depreciation rate. Suppose that there is an expansionary monetary policy shock that drives down the real interest rate, with the maximal impact occurring

[3] For example, in the finance literature, see Eichenbaum and Hansen (1990), Constantinides (1990), and Boldrin, Christiano, and Fisher (2001). In the growth literature, see Carroll, Overland, and Weil (1997, 2000). In the psychology literature, see Gremel et al. (2016).

contemporaneously, as in the data. Absent adjustment costs, arbitrage then requires that the marginal product of capital follow a pattern identical to the real interest rate. For that to happen, both the capital stock and investment must have exactly the opposite pattern to the marginal product of capital. With the biggest surge in investment occurring in the period of the monetary policy shock, the simple model cannot reproduce the hump-shape pattern in Figure 1. When it is costly to adjust the rate of investment, households choose to raise investment slowly over time, generating a hump-shape response pattern as in Figure 1.

Lucca (2006) and Matsuyama (1984) provide interesting theoretical foundations for the investment adjustment cost in Christiano, Eichenbaum, and Evans (2005). In addition, there is substantial empirical evidence in support of the specification (Eberly, Rebelo, and Vincent 2012; Matsuyama 1984).

An important alternative specification of adjustment costs penalizes changes in the capital stock. This specification has a long history in macroeconomics, going back at least to Lucas and Prescott (1971). Christiano, Eichenbaum, and Evans (2005) show that with this type of adjustment cost, investment jumps after an expansionary monetary policy shock and then converges monotonically back to its pre-shock level from above. This response pattern is inconsistent with the vector autoregression evidence.

Nominal Rigidities. In contrast to real business cycle models, goods and labor markets in Christiano, Eichenbaum, and Evans (2005) are not perfectly competitive. This departure is necessary to allow for sticky prices and sticky nominal wages—if a price or wage is sticky, someone has to set it. In this model, nominal rigidities arise from Calvo (1983)-style frictions. In particular, firms and households can change prices or wages with some exogenous probability. In addition, they must satisfy whatever demand materializes at those prices and wages.

Calvo-style frictions make sense only in environments where inflation is moderate. Even in moderate inflation environments, Calvo-style frictions have implications that are inconsistent with aspects of micro data (for example, Nakamura and Steinsson 2008; Eichenbaum, Jaimovich, and Rebelo 2011). Still, the continued use of this assumption reflects two factors. First, Calvo-style frictions allow models to capture, in an elegant and tractable manner, what many researchers believe is an essential feature of business cycles; for moderate inflation economies, firms and labor suppliers typically respond to variations in demand by varying quantities rather than prices. Second, authors like Eichenbaum, Jaimovich, and Rebelo (2011) argue that, for moderate inflation economies, the Calvo model provides a good approximation to more plausible models in which firms face costs of changing their pricing strategies.

Acyclical Marginal Costs. Christiano, Eichenbaum, and Evans (2005) build features into the model which ensure that firms' marginal costs are nearly acyclical. They do so for three reasons. First, there is substantial empirical evidence in favor of this view (for example, Anderson, Rebelo, and Wong 2018). Second, the more acyclical is marginal cost, the more plausible is the assumption that firms satisfy demand. Third, as in standard New Keynesian models, inflation is an increasing function of current and expected future marginal costs. Thus, relatively acyclical marginal costs are critical for dampening movements in the inflation rate.

The model in Christiano, Eichenbaum, and Evans (2005) incorporates two mechanisms to ensure that marginal costs are relatively acyclical: the sticky nominal wage assumption mentioned above; and the rate at which capital is utilized can be varied in response to shocks.

Quantitative Properties. To illustrate the model's quantitative properties, we work with the variant of the model of Christiano, Eichenbaum, and Evans (2005) estimated in Christiano, Eichenbaum, and Trabandt (2016). We re-estimated the model using a Bayesian procedure that treats the impulse responses to a monetary policy shock based on vector autoregressions as data. The online Appendix to this paper provides details about the prior and posterior distributions of model parameters. Here we highlight some of the key estimated parameters. The posterior mode estimates imply that firms change prices on average once every 2.3 quarters; the household changes nominal wages about once a year; past consumption enters with a coefficient of 0.75 in the household's utility function; and the elasticity of investment with respect to a one percent temporary increase in the current price of installed capital is equal to 0.16.

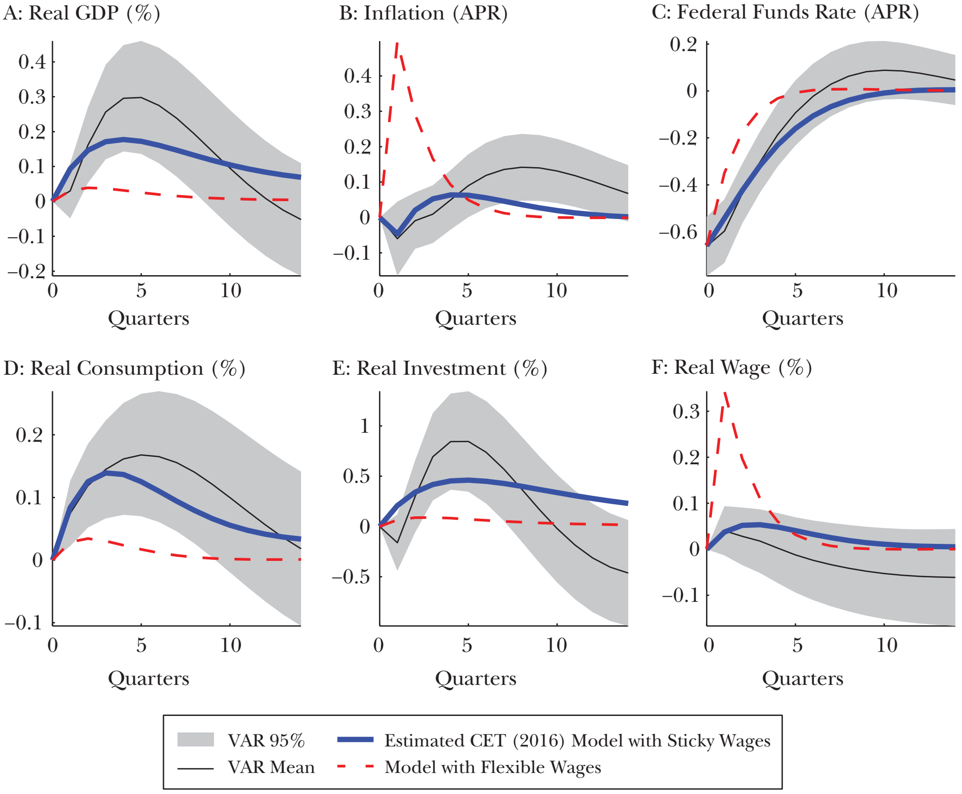
The thin solid line in the panels of Figure 2 is the impulse response function estimate reproduced from Figure 1. The grey area depicts the 95 percent confidence intervals associated with that estimate. The thicker solid line depicts the impulse response function of the estimated DSGE model to a monetary policy shock, calculated using the mode of the posterior distribution of the model's parameters.

Four key features of the results are worth noting. First, the model succeeds in accounting for the hump-shape rise in consumption, investment, and real GDP after a policy-induced fall in the federal funds rate. Second, the model succeeds in accounting for the small rise in inflation after the shock. Third, the model has the property that real wages are essentially unaffected by the policy shock. Finally, the model has the anti-Fisherian property that the nominal interest rate and inflation move in the opposite direction after a transitory monetary policy shock.

We emphasize that the model's properties depend critically on sticky wages. The dashed line in Figure 2 depicts the model's implications if we recalculate the impulse responses assuming that nominal wages are fully flexible (holding other model parameters fixed at the mode of the posterior distribution). Note that the model's performance deteriorates drastically. In Christiano, Eichenbaum, and Evans (2005), sticky wages are sticky by assumption. In Christiano, Eichenbaum, and Trabandt (2016), we show that wage stickiness arises endogenously in a version of the Christiano, Eichenbaum, and Evans (2005) model that has labor market search and matching frictions. The key feature of the model is that workers and firms bargain in a way that reduces the sensitivity of the wage to macroeconomic aggregates. One advantage of endogenously generating sticky wages in this way is we can analyze the aggregate effects of various policies like unemployment insurance. Finally, we note that habit formation and investment adjustment costs are critical to the model's success. Absent those features, it would be very difficult to generate hump-shaped responses with reasonable degrees of nominal rigidities.

Figure 2

Impulse Responses to a Monetary Policy Shock: Vector Autoregression (VAR) versus Model



Source: Authors.

Note: The thin solid lines in Figure 2 are the impulse response function estimates reproduced from Figure 1. All data are expressed in deviations from what would have happened in the absence of the shock. The units are given in the titles of the subplots. Percent means percent deviation from unshocked path. The grey area depicts the 95 percent confidence intervals associated with those estimates. The thicker solid line depicts the impulse response function of the dynamic stochastic general equilibrium model to a monetary policy shock, calculated using the mode of the posterior distribution of the model's parameters. The dashed line depicts the model's implications if we recalculate the impulse responses assuming that nominal wages are fully flexible (holding other model parameters fixed at the mode of the posterior distribution).

How Dynamic Stochastic General Equilibrium Models Are Estimated and Evaluated

Prior to the financial crisis, researchers generally worked with log-linear approximations to the equilibria of dynamic stochastic general equilibrium models. There were three reasons for this choice. First, for the models being considered and for the size of shocks that seemed relevant for the postwar US data, linear approximations are very accurate (for discussion, see the papers in Taylor and Uhlig 1990). Second, linear approximations allow researchers to exploit the large array of tools for forecasting, filtering, and estimation provided in the literature on linear time series analysis. Third, it was simply not computationally feasible to solve and estimate large, nonlinear DSGE models. The technological constraints were real and binding.

Researchers choose values for the key parameters of their models using a variety of strategies. In some cases, researchers choose parameter values to match unconditional model and data moments, or they reference findings in the empirical micro literature. This procedure is called calibration and does not use formal sampling theory. Calibration was the default procedure in the early real business cycle literature, and it is also sometimes used in the dynamic stochastic general equilibrium literature. Most of the modern DSGE literature conducts inference about parameter values and model fit using one of two strategies that make use of formal econometric sampling theory: limited information and full information.

The limited information strategy does not exploit all of the model's implications for moments of the data. One variant of this strategy minimizes the distance between a subset of model-implied second moments and their analogs in the data. A more influential variant of this first strategy estimates parameters by minimizing the distance between model and data impulse responses to economic shocks. Examples of this impulse response matching approach include Christiano, Eichenbaum, and Evans (2005), Altig, Christiano, Eichenbaum, and Linde (2011), Iacoviello (2005), and Rotemberg and Woodford (1991).

One way to estimate the data impulse response functions is based on partially identified vector autoregressions. Another variant of this strategy, sometimes referred to as the method of external instruments, involves using historical or narrative methods to obtain instruments for the underlying shocks (Mertens and Ravn 2013). Finally, researchers have exploited movements in asset prices immediately after central bank policy announcements to identify monetary policy shocks and their consequences. This approach is referred to as high frequency identification (for example, early contributions include Kuttner 2001; Gürkaynak, Sack, and Swanson 2005).

The initial limited information applications in the DSGE literature used generalized method of moments estimators and classical sampling theory (Hansen 1982). Building on the work of Chernozhukov and Hong (2003), Christiano, Trabandt, and Walentin (2010) showed how the Bayesian approach can be applied in limited information contexts. A critical advantage of the Bayesian approach is that one can formally and transparently bring to bear information from a variety of sources on what constitutes "reasonable" values for model parameters. Suppose, for example, that one could only match the dynamic response to a monetary policy shock for model parameter values that firms change their prices on average every two years. This implication is strongly at variance with evidence from micro data. In the Bayesian approach, the analyst would impose priors that sharply penalize such parameter values, so that those parameter values would be assigned low probabilities in the analyst's posterior distribution. Best practice compares priors and posteriors for model parameters. This comparison allows the analyst to make clear the role of priors and the data in generating the results.

At a deeper level, micro data influences, in a critical but slow-moving manner, the class of models with which we work. Our discussion of the demise of the pure real business cycle model is one illustration of this process. The models of financial frictions and heterogeneous agents discussed below are an additional illustration of how DSGE models evolve over time in response to micro data.

The other strategy for estimating dynamic stochastic general equilibrium models involves full-information methods. In many applications, the data used for estimation is relatively uninformative about the value of some of the parameters in DSGE models (Canova and Sala 2009). A natural way to deal with this fact is to bring other information to bear on the analysis. Bayesian priors are a vehicle for doing exactly that, which is an important reason why the Bayesian approach has been very influential in full-information applications. Starting from Smets and Wouters (2003), a large econometric literature has expanded the Bayesian toolkit to include better ways to conduct inference about model parameters and to analyze model fit. For a recent survey, see Fernández-Villaverde, Rubio-Ramierez, and Schorfheide (2016).

Why Didn't DSGE Models Predict the Financial Crisis?

Pre-crisis dynamic stochastic general equilibrium models did not predict the increasing vulnerability of the US economy to a financial crisis. They have also been criticized for not placing more emphasis on financial frictions. Here, we give our perspective on these failures.

The debate about the causes of the financial crisis is ongoing. Our view, shared by Bernanke (2009) and many others, is that the financial crisis was precipitated by a rollover crisis in a very large and highly levered shadow-banking sector that relied on short-term debt to fund long-term assets. By shadow banks, we mean financial institutions not covered by the protective umbrella of the Federal Reserve and Federal Deposit Insurance Corporation (for further discussion, see Bernanke 2010).

This rollover crisis was triggered by a set of developments in the housing sector. US housing prices began to rise rapidly in the 1990s. The S&P/Case-Shiller U.S. National Home Price Index rose by a factor of roughly 2.5 between 1991 and 2006. The precise role played by expectations, the subprime market, declining lending standards in mortgage markets, and overly loose monetary policy is not critical for our purposes. What is critical is that housing prices began to decline in mid-2006, causing a fall in the value of the assets of shadow banks that had heavily invested in mortgage-backed securities. The Fed's willingness to provide a safety net for the shadow banking system was at best implicit, creating the conditions under which a rollover crisis was possible. In fact, a rollover crisis did occur and shadow banks had to sell their asset-backed securities at fire-sale prices, precipitating the financial crisis and the Great Recession.

Against this background, we turn to the two criticisms of dynamic stochastic general equilibrium models mentioned above. The first criticism, namely the failure to signal the increasing vulnerability of the US economy to a financial crisis, is correct. The failure reflected a broader failure of the economics community.

The overwhelming majority of academics, regulators, and practitioners did not realize that a small shadow-banking system had metastasized into a massive, poorly regulated, Wild-West-like sector that was not protected by deposit insurance or lender-of-last-resort backstops.

The second criticism of dynamic stochastic general equilibrium models was that they did not sufficiently emphasize financial frictions. In practice, modelers have to make choices about which frictions to emphasize. One reason why modelers did not emphasize financial frictions in DSGE models is that until the Great Recession, postwar recessions in the United States and western Europe did not seem closely tied to disturbances in financial markets. The savings and loans crisis in the US economy in the late 1980s and early 1990s was a localized affair that did not grow into anything like the Great Recession. Similarly, the stock market meltdown in 1987 and the bursting of the tech bubble in 2001 only had minor effects on aggregate economic activity.

At the same time, the financial frictions that were included in dynamic stochastic general equilibrium models did not seem to have very big effects. Consider, for example, Bernanke, Gertler, and Gilchrist's (1999) model, which is arguably the most influential pre-crisis DSGE model with financial frictions. The financial accelerator in that model has only a modest quantitative effect on the way the model economy responds to shocks (see for example, Lindé, Smets, and Wouters 2016). In the same spirit, Kocherlakota (2000) argues that models with credit constraints (of the type in Kiyotaki and Moore 1997) have only negligible effects on dynamic responses to shocks. Finally, Brzoza-Brzezina and Kolasa (2013) compare the empirical performance of the standard New Keynesian DSGE model with variants that incorporate Kiyotaki and Moore (1997) and Bernanke, Gertler, and Gilchrist (1999)-type constraints. Their key finding is that neither model substantially improves on the performance of the benchmark model, either in terms of marginal likelihoods or impulse response functions. Thus, guided by the postwar data from the United States and western Europe and experience with existing models of financial frictions, DSGE modelers emphasized other frictions.

After the Storm

Given the data-driven nature of the dynamic stochastic general equilibrium enterprise, it is not surprising that the financial crisis and its aftermath had an enormous impact on dynamic stochastic general equilibrium models. In this section, we discuss the major strands of work in post–financial crisis DSGE models.

Financial Frictions

The literature on financial frictions can loosely be divided between papers that focus on frictions originating inside financial institutions and those that arise from the characteristics of the people who borrow from financial institutions. Theories of bank runs and rollover crisis focus on the first class of frictions. Theories of collateral constrained borrowers focus on the second class of frictions. This is not the place for a systematic review, but here we discuss some examples of each.

Frictions that Originate Inside Financial Institutions. Motivated by events associated with the financial crisis, Gertler and Kiyotaki (2015) and Gertler, Kiyotaki, and Prestipino (2016) develop a dynamic stochastic general equilibrium model of a rollover crisis in the shadow banking sector, which triggers fire sales. The resulting decline in asset values tightens balance sheet constraints in the rest of the financial sector and throughout the economy.[4]

In the Gertler and Kiyotaki (2015) model, shadow banks finance the purchase of long-term assets by issuing short-term (one-period) debt. Banks have two ways to deal with short-term debt that is coming due. The first is to issue new short-term debt (that is, "rolling over the debt"). The second is to sell assets. The creditor's only decision is whether to buy new short-term debt. There is nothing the creditor can do to affect payments received on past short-term debt.[5]

There is always an equilibrium in the Gertler and Kiyotaki (2015) model in which shadow banks can roll over the short-term debt without incident. But, there can also be an equilibrium in which each creditor chooses not to roll over the debt. Suppose that an individual creditor believes that other creditors won't extend new credit to banks. In that case, there will be a system-wide failure of the banks, as attempts to pay off bank debt lead to fire sales of assets that wipe out bank equity. The individual creditor would prefer to buy assets at fire sale prices rather than extend credit to a bank that has zero net worth. With every potential creditor thinking this way, it is a Nash equilibrium for each creditor not to purchase new liabilities from banks. Such an equilibrium is referred to as a rollover crisis.

A rollover crisis leads to fire sales because, with all banks selling, the only potential buyers are other agents who have little experience evaluating the banks' assets. In this state of the world, agency problems associated with asymmetric information become important.[6]

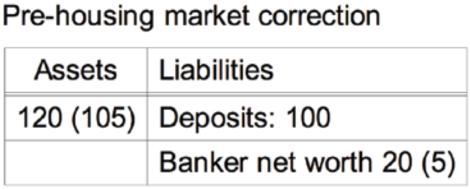
As part of the specification of the model, Gertler and Kiyotaki (2015) assume that the probability of a rollover crisis is proportional to the losses that depositors would experience in the event that a rollover crisis occurs. Thus, if bank creditors think that banks' net worth would be positive in a crisis, then a rollover crisis is impossible. However, if banks' net worth is negative in this scenario, then a rollover crisis can occur.

We use this model to illustrate how a relatively small shock can trigger a system-wide rollover crisis in the shadow banking system. To this end, consider the following illustrative example, which captures in a highly stylized way the key features of the shadow-banking system before (left-side table below) and after (right- side table below) the crisis.

[4] The key theoretical antecedent is the bank run model of Diamond and Dybvig (1983) and the sovereign debt rollover crisis model of Cole and Kehoe (2000).

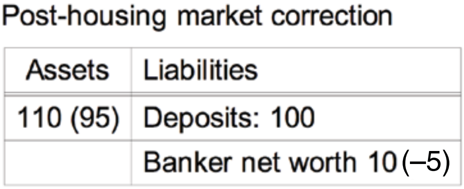
[5] Unlike in the classic bank run model of Diamond and Dybvig (1983), there is no reason to impose a sequential debt service constraint.

[6] Gertler and Kiyotaki (2015) capture these agency problems by assuming that the buyers of long-term assets during a rollover crisis are relatively inefficient at managing the assets.



Pre-housing market correction

Assets	Liabilities
120 (105)	Deposits: 100
	Banker net worth 20 (5)



Post-housing market correction

Assets	Liabilities
110 (95)	Deposits: 100
	Banker net worth 10 (–5)

In the left-side table, the shadow banks' assets and liabilities are 120 and 100, respectively—so their net worth is positive. The numbers in parentheses show the value of the assets and net worth of the shadow banks if there were to be a rollover crisis and fire-sale of assets. Since net worth remains positive, the Gertler and Kiyotaki (2015) analysis implies that a rollover crisis cannot occur.

Now imagine that the assets of the shadow banks decline because of a small shift in fundamentals. Here, we have in mind the events associated with the decline in housing prices that began in the summer of 2006. The right-side table is the analog of the left side, taking into account the lower value of the shadow banks' assets. In the example, the market value of assets has fallen by 10, from 120 to 110. In the absence of a rollover crisis, the system is solvent. However, the value of the assets in the case of a rollover crisis is 95 and the net worth of the bank is negative in that scenario. Thus, a relatively small change in asset values could lead to a severe financial crisis.

The example illustrates two important potential uses of DSGE models. First, an estimated DSGE model can be used to calculate the probability of a rollover crisis, conditional on the state of the economy. In principle, one could estimate this probability function using reduced form methods. However, because financial crises are rare events, estimates emerging from reduced form methods would have enormous sampling uncertainty. Because of its general equilibrium structure, an empirically plausible DSGE model would address the sampling uncertainty problem by making use of a wider array of information drawn from non-crisis times to assess the probability of a financial crisis.

Second, DSGE models can potentially be used to design policies that address financial crises. While we think that existing DSGE models of financial crisis such as Gertler and Kiyotaki (2015) yield valuable insights, these models are clearly still in their infancy. For example, the model assumes that people know what can happen in a crisis, together with the associated probabilities. This seems implausible, given the fact that a full-blown crisis happens two or three times per century in developed economy like the United States. It seems safe to conjecture that factors such as aversion to "Knightian uncertainty" play an important role in driving fire sales in a crisis (see for example, Caballero and Krishnamurthy 2008). Still, research on various types of crises is proceeding at a rapid pace, and we expect to see substantial improvements in DSGE models on the subject (for examples, see Bianchi, Hatchondo, and Martinez 2016 and the references therein).

Frictions Associated with the People that Borrow from Financial Institutions. We now turn to the second set of financial frictions. One of the themes of this paper is that data analysis lies at the heart of the DSGE project. Elsewhere, we have stressed the importance of microeconomic data. Here, we also stress the role of financial data as a source of information about the sources of economic fluctuations. Using an estimated DSGE model, Christiano, Motto, and Rostagno (2014) argue that the dominant source of US business cycle fluctuations are disturbances in the riskiness of individual firms (what they call "risk shocks"). A motivation for their analysis is that in recessions, firms pay a premium to borrow money, above the rate at which a risk-free entity like the US government borrows. They interpret this premium as, in effect, reflecting the view of lenders that firms represent a riskier bet. Christiano, Motto, and Rostagno (2014) estimate their DSGE model using a large number of macroeconomic and financial variables and conclude that fluctuations in risk can account for the bulk of GDP fluctuations.

To understand the intuition behind the model, consider a recession that is triggered by an increase in the riskiness of firms.[7] As the cost of borrowing rises, firms borrow less and demand less capital. This decline induces a fall in both the quantity and price of capital. In the presence of nominal rigidities and a Taylor rule for monetary policy, the decline in investment leads to an economy-wide recession, including a fall in consumption and a rise in firm bankruptcies. With the decline in aggregate demand, inflation falls. Significantly, the risk shock leads to an increase in the cross-sectional dispersion of the rate of return on firm equity. Moreover, the recession is also associated with a fall in the stock market, driven primarily by capital losses associated with the fall in the price of capital. All these effects are observed in a typical recession.[8] This is why Christiano, Motto, and Rostagno's (2014) estimation procedure attributes 60 percent of the variance of US business cycles to risk shocks.

The dynamic effects of risk shocks in the Christiano, Motto, and Rostagno (2014) model resemble business cycles so well that many of the standard shocks that appear in previous business cycle models are rendered unimportant in the empirical analysis. For example, Christiano et al. (2014) find that aggregate shocks to the technology for producing new capital account for only 13 percent of the business cycle variation in GDP. This contrasts sharply with the results in Justiniano, Primiceri, and Tambalotti (2010), who argue that this shock accounts for roughly 50 percent of business cycle variation of GDP. The critical difference is that Christiano, Motto, and Rostagno (2014) include financial data like the stock market in their analysis. Shocks to the supply of capital give rise to countercyclical movements in the stock market, so they cannot be the prime source of business cycles.

Financial frictions have also been incorporated into a growing literature that introduces the housing market into DSGE models. One part of this literature focuses on the implications of housing prices for households' capacity to borrow (see for example, Iacoviello and Neri 2010; Berger, Guerrieri, Lorenzoni, and Vavra forthcoming). Another part focuses on the implications of land and housing prices on firms' capacity to borrow (Liu, Wang, and Zha 2013).

[7] In Christiano, Motto, and Rostagno (2014), a rise in risk corresponds to an increase in the variance of a firm-specific shock to technology. Absent financial frictions, such a shock would have no impact on aggregate output. A rise in the variance would lead to bigger-sized shocks at the firm level but the average across firms is only a function of the mean (law of large numbers).

[8] To our knowledge, the first paper to articulate the idea that a positive shock to idiosyncratic risk could produce effects that resemble a recession is Williamson (1987).

Zero Lower Bound and Other Nonlinearities

The financial crisis and its aftermath was associated with two important nonlinear phenomena. The first phenomenon was the rollover crisis in the shadow-banking sector discussed above. The Gertler and Kiyotaki (2015) model illustrates the type of nonlinear model required to analyze this type of crisis. The second phenomenon was that the nominal interest rate hit the zero-lower bound in December 2008. An earlier theoretical literature associated with Krugman (1998), Benhabib, Schmitt-Grohé, and Uribe (2001), and Eggertsson and Woodford (2003) had analyzed the implications of the zero-lower bound for the macroeconomy. Building on this literature, DSGE modelers quickly incorporated the zero-lower bound into their models and analyzed its implications.

In what follows, we discuss how nonlinear DSGE models have been used to assess which shocks triggered the financial crisis and what propagated their effects over time. We focus on three papers to give the reader a flavor of this literature. We then review some of the policy advice relating to fiscal policy and forward guidance that emerges from recent DSGE models that incorporate a zero lower bound.

The Causes of the Crisis and Slow Recovery. Several DSGE models provide a quantitatively plausible description of the behavior of major economic aggregates during the Great Recession when the zero lower bound was a binding constraint. In Christiano, Eichenbaum, and Trabandt (2015), we analyze the post-crisis period taking into account that the zero lower bound was binding. In addition, we take into account the forward guidance of the Federal Reserve Open Market Committee about future monetary policy. This guidance was highly nonlinear in nature: it involved a regime switch depending on the realization of endogenous variables (like the unemployment rate). We argue that the bulk of movements in aggregate real economic activity during the Great Recession was due to financial frictions interacting with the zero lower bound. Our analysis also indicates that the observed fall in total factor productivity and the rise in the cost of working capital played important roles in accounting for the surprisingly small drop in inflation after the financial crisis.

Lindé and Trabandt (2018) argue that nonlinearities in price and wage-setting are an alternative reason for the small decline in inflation during the Great Recession. In particular, they assume that the elasticity of demand of a goods-producing firm is increasing in its relative price along the lines proposed in Kimball (1995). So, during a recession when marginal costs are falling, firms that can change their prices have less of an incentive to do so relative to the case in which the elasticity of demand is constant. They show that this effect is quantitatively important in the standard nonlinear New Keynesian DSGE model.

Gust, Herbst, López-Salido, and Smith (2017) estimate a fully nonlinear DSGE model with an occasionally binding zero lower bound. Nonlinearities in the model play an important role for inference about the source and propagation of shocks. According to their analysis, shocks to the demand for risk-free bonds and, to a lesser extent, the marginal efficiency of investment proxying for financial frictions, played a critical role in the crisis and its aftermath.

Critically, the above papers include both financial frictions and nominal rigidities. A model of the crisis and its aftermath that didn't have financial frictions would not be plausible. At the same time, a model that included financial frictions but didn't allow for nominal rigidities would have difficulty accounting for the broad-based decline across all sectors of the economy. For example, such a model would predict a boom in sectors of the economy that are less dependent on the financial sector.

The fact that DSGE models with nominal rigidities and financial frictions can provide quantitatively plausible accounts of the financial crisis and the Great Recession makes them obvious frameworks within which to analyze alternative policies. We begin with a discussion of fiscal policy.

Fiscal Policy. In standard DSGE models, an increase in government spending triggers a rise in output and inflation. When monetary policy is conducted according to a standard Taylor rule, a rise in inflation triggers a rise in the real interest rate. Other things equal, the policy-induced rise in the real interest rate lowers investment and consumption demand. As a result, the government spending multiplier is typically less than one in these models. But when the zero lower bound binds, the rise in inflation associated with an increase in government spending does not trigger a rise in the real interest rate. With the nominal interest rate stuck at zero, a rise in inflation lowers the real interest rate, crowding consumption and investment in, rather than out. This raises the quantitative question: how does a binding zero lower bound constraint on the nominal interest rate affect the size of the government spending multiplier?

Christiano, Eichenbaum, and Rebelo (2011) address this question in a DSGE model, assuming all taxes are lump sum. A basic principle that emerges from their analysis is that the multiplier is larger the more binding is the zero lower bound. They measure how binding the zero lower bound is by how much a policymaker would like to lower the nominal interest rate below zero if it were possible to do so. For their preferred specification, the multiplier is much larger than one. When the zero lower bound is not binding, then the multiplier would be substantially below one.

Erceg and Lindé (2014) examine (among other things) the impact of distortionary taxation on the magnitude of the government spending multiplier in the zero lower bound. They find that the results based on lump-sum taxation are robust relative to the situation in which distortionary taxes are raised gradually to pay for the increase in government spending.

At this point, a large literature now studies the fiscal multiplier when the zero lower bound can bind using DSGE models that allow for financial frictions, open-economy considerations and liquidity constrained consumers. Such models are playing an important role in the debate among academics and policymakers about whether and how fiscal policy should be used to fight recessions. We offer two examples. First, Coenen et al. (2012) analyze the impact of different fiscal stimulus shocks in several DSGE models that are used by policy-making institutions. Second, Blanchard, Erceg, and Lindé (2017) analyze the effects of a fiscal expansion by the core euro area economies on the periphery euro area economies. Finally, we note that the early papers on the size of the government spending multiplier use log-linearized versions of DSGE models. For example, Christiano, Eichenbaum, and Rebelo (2011) work with a linearized version of their model, while in Christiano, Eichenbaum, and Trabandt (2015), we work with a nonlinear version of the model. Significantly, there is now a literature that assesses the sensitivity of multiplier calculations to linear versus nonlinear solutions (for example, Christiano and Eichenbaum 2012; Boneva, Braun, and Waki 2016; Christiano, Eichenbaum, and Johannsen 2017; Lindé and Trabandt forthcoming).

Forward Guidance. When the zero lower bound constraint on the nominal interest rate became binding, conventional monetary policy (that is, lowering short-term interest rates) was no longer possible. Monetary policymakers considered a variety of alternatives. Here, we focus on forward guidance as a policy option analyzed by Eggertsson and Woodford (2003) and Woodford (2012) in New Keynesian models. By forward guidance, we mean that the monetary policymaker promises to keep the policy interest rate lower for longer than the monetary rule would otherwise suggest.

As documented in Carlstrom, Fuerst, and Paustian (2015), forward guidance is implausibly powerful in standard DSGE models like Christiano, Eichenbaum, and Evans (2005). Del Negro, Giannoni, and Patterson (2012) refer to this phenomenon as the "forward guidance puzzle." This puzzle has fueled an active debate. Carlstrom, Fuerst, and Paustian (2015) and Kiley (2016) show that the magnitude of the forward guidance puzzle is substantially reduced in a sticky information (as opposed to a sticky price) model. Other responses to the forward guidance puzzle involve more fundamental changes, such as abandoning the representative agent framework, which is discussed in the next subsection. More radical responses involve abandoning strong forms of rational expectations (see for example Gabaix 2016; Woodford 2018; Angeletos and Lian forthcoming).

Heterogeneous Agent Models

In the standard New Keynesian model, the primary channel by which monetary-policy-induced interest rate changes affect consumption is by causing the representative household to reallocate consumption over time. However, there is a great deal of empirical micro evidence against the importance of this reallocation channel, in part because many households face binding borrowing constraints.

Motivated by these observations, macroeconomists are exploring DSGE models where heterogeneous consumers face idiosyncratic shocks and binding borrowing constraints.[9] Kaplan, Moll, and Violante (2018) and McKay, Nakamura, and Steinsson (2016) are useful starting points that convey the flavor of the literature. Both of these papers present DSGE models in which households have uninsurable, idiosyncratic income risk, and in which many households face borrowing constraints.[10]

The literature on DSGE models with heterogeneous agents is still young, but it has already yielded important insights into important policy issues like the impact of forward guidance (McKay, Nakamura, and Steinsson 2016; Farhi and Werning 2017). The literature has also led to a richer understanding of how monetary policy actions affect the economy. In Kaplan, Moll, and Violante (2018), for example, a monetary policy action initially affects the small set of households who

[9] There is also important work allowing for firm heterogeneity in DSGE models (for examples, see Gilchrist, Schoenle, Sim, and Zakrajšek 2017; Ottonello and Winberry 2017).

[10] Important earlier papers in this literature include Oh and Reis (2012), Guerrieri and Lorenzoni (2017), McKay and Reis (2016), Gornemann, Kuester, and Nakajima (2016), and Auclert (2017).

actively intertemporally adjust spending in response to an interest rate change. However, most of the impact occurs through a multiplier-type process that occurs as other firms and households adjust their spending in response to the change in demand by the "intertemporal adjusters." This area of research typifies the cutting edge of DSGE models: the key features are motivated by micro data, and the implications (say, for the multiplier-type process) are assessed using both micro and macro data.

How Are DSGE Models Used in Policy Institutions?

As a case study of how DSGE models are used in policy institutions, we focus on the Board of Governors of the Federal Reserve System. We are guided in our discussion by Stanley Fischer's (2017) description of the policy-making process at the Federal Reserve Board.

Before the Federal Open Market Committee meets to make policy decisions, all participants are given copies of the so-called Tealbook, which includes parts A and B.[11] Tealbook A contains a summary and analysis of recent economic and financial developments in the United States and foreign economies, as well as the staff's economic forecast. The staff also provides model-based simulations of a number of alternative scenarios highlighting upside and downside risks. Examples of such scenarios include a decline in the price of oil, a rise in the value of the dollar or wage growth that is stronger than the one built into the baseline forecast. These scenarios are generated using one or more of the Board's macroeconomic models, including the DSGE models, SIGMA, and EDO.[12] Tealbook A also contains estimates of future outcomes in which the Federal Reserve Board uses alternative monetary policy rules as well model-based estimates of optimal monetary policy. According to Fischer (2017), DSGE models play a central, though not exclusive, role in this process.

Tealbook B provides an analysis of specific policy options. According to Fischer (2017), "Typically, there are three policy alternatives—A, B, and C—ranging from dovish to hawkish, with a centrist one in between." DSGE models, along with other approaches, are used to generate the quantitative implications of the specific policy alternatives considered.

The Federal Reserve System is not the only policy institution that uses DSGE models. For example, the European Central Bank, the International Monetary Fund, the Bank of Israel, the Czech National Bank, the Sveriges Riksbank, the Bank of Canada, and the Swiss National Bank all use such models in their policy process.[13]

[11] The Tealbooks are available with a five-year lag at https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm.

[12] For a discussion of the SIGMA and EDO models, see Erceg, Guerrieri, and Gust (2006) and, at the Federal Reserve website, https://www.federalreserve.gov/econres/edo-models-about.htm.

[13] For a review of the dynamic stochastic general equilibrium models used in the policy process at the European Central Bank, see Smets, Christoffel, Coenen, Motto, and Rostagno (2010). Carabenciov et al. (2013) and Freedman, Kumhof, Laxton, Muir, and Mursual (2009) describe global DSGE models used for policy analysis at the International Monetary Fund (IMF), while Benes, Kumhof, and Laxton (2014) describe

We just argued that DSGE models are used to run policy simulations in various policy institutions. The results of those simulations are useful to the extent that the models are empirically plausible. One important way to assess the plausibility of a model is to consider its real time forecasting performance. Cai, Del Negro, Giannoni, Gupta, Li, and Moszkowski (2018) compare real-time forecasts of the New York Fed DSGE model with those of various private forecasters and with the median forecasts of the Federal Open Market Committee members. The DSGE model that they consider is a variant of Christiano, Motto, and Rostagno (2014) that allows for shocks to the demand for government bonds. Cai et al. find that the model-based real time forecasts of inflation and output growth are comparable to those of private forecasters. Strikingly, the New York Fed DSGE model does a better job at forecasting the slow recovery than the Federal Open Market Committee, at least as judged by the root mean square errors of their median forecasts. Cai et al. argue that financial frictions play a critical role in allowing the model to anticipate the slow growth in output after the financial crisis.

In sum, dynamic stochastic general equilibrium models play an important role in the policymaking process. To be clear: they do not substitute for judgment, nor should they. But policymakers have voted with their collective feet on the usefulness of DSGE models. In this sense, DSGE models are meeting the market test.

A Brief Response to Some Critiques

Here, we briefly respond to some recent critiques of dynamic stochastic general equilibrium models. We focus on Stiglitz (2017) because his critique is both well known and representative of some common criticisms.

Econometric Methods

Stiglitz (2017, p. 1), citing what he refers to as Korinek (2017)'s "devastating critique" of DSGE practitioners, claims: "Standard statistical standards are shunted aside [by DSGE modelers]." As evidence, he writes: "[T]he time series employed are typically detrended using methods such as the HP [Hodrick–Prescott] filter to focus the analysis on stationary fluctuations at business cycle frequencies. Although this is useful in some applications, it risks throwing the baby out with the bathwater as many important macroeconomic phenomena are non-stationary or occur at lower frequencies" (p. 3). But this criticism is simply

MAPMOD, a DSGE model used at the IMF for the analysis of macroprudential policies. Clinton, Hlédik, Holub, Laxton, and Wang (2017) describe the role of DSGE models in policy analysis at the Czech National Bank, and Adolfson, Laséen, Christiano, Trabandt, and Walentin (2013) describe the RAMSES II DSGE model used for policy analysis at the Sveriges Riksbank. Argov et al. (2012) describe the DSGE model used for policy analysis at the Bank of Israel, Dorich, Johnston, Mendes, Murchison, and Zhang (2013) describe ToTEM, the DSGE model used at the Bank of Canada for policy analysis, and Alpanda, Cateau, and Meh (2014) describe MP2, the DSGE model used at the Bank of Canada to analyze macroprudential policies. Rudolf and Zurlinden (2014) and Gerdrup, Kravik, Paulsen, and Robstad (2017) describe the DSGE model used at the Swiss National Bank and the Norges bank, respectively, for policy analysis.

incorrect. The vast bulk of the modern DSGE literature does not estimate models using HP filtered data. Moreover, DSGE models of endogenous growth provide a particularly stark counterexample to the claim that this approach is limited to the analysis of stationary fluctuations at business cycle frequencies. Notably, neither Stiglitz nor Korinek offer any constructive advice on how to address the difficult problem of dealing with nonstationary data. In sharp contrast, the DSGE literature struggles mightily with this problem and adopts different strategies for modeling non-stationarity in the data. As one example, see Comin and Gertler (2006)'s analysis of medium-term business cycles.

Stiglitz (2017) then claims (pp. 3–4) "for given detrended time series, the set of moments chosen to evaluate the model and compare it to the data is largely arbitrary—there is no strong scientific basis for one particular set of moments over another … [F]or a given set of moments, there is no well-defined statistic to measure the goodness of fit of a DSGE model or to establish what constitutes an improvement in such a framework." This criticism might have been appropriate in the 1980s. But, it simply does not apply to modern analyses, which use full information maximimum likelihood or generalized method of moments.

Financial Frictions

Stiglitz (2017) asserts that pre-crisis DSGE models did not allow for financial frictions or liquidity-constrained consumers. This claim is incorrect. As one example, Galí, López-Salido, and Vallés (2007) investigate the implications of the assumption that some consumers are liquidity constrained. Specifically, they assume that a fraction of households cannot borrow at all. They then assess how this change affects the implications of DSGE models for the effects of a shock to government consumption. Not surprisingly, they find that liquidity constraints substantially magnify the impact of government spending on GDP. Looking back further, Carlstrom and Fuerst (1997) and Bernanke, Gertler, and Gilchrist (1999) develop DSGE models that incorporate credit market frictions that give rise to a "financial accelerator" in which credit markets work to amplify and propagate shocks to the macroeconomy.

In other examples, Christiano, Motto, and Rostagno (2003) add several features to the model of Christiano, Eichenbaum, and Evans (2005) to allow for richer financial markets. They incorporate the fractional reserve banking model developed by Chari, Christiano, and Eichenbaum (1995). They allow for financial frictions as modeled by Bernanke, Gertler, and Gilchrist (1999) and Williamson (1987). In addition they assume that agents can only borrow using nominal non-state-contingent debt, so that the model incorporates the Fisherian debt deflation channel. Finally, Iacoviello (2005) develops and estimates a DSGE model with nominal loans and collateral constraints tied to housing values. This paper is an important antecedent to the large post-crisis DSGE literature on the aggregate implications of housing market booms and busts.

Stiglitz (2017) also asserts that DSGE models abstract from interest rate spreads. He writes (p. 10): "... in standard [DSGE] models ... all that matters is that somehow the central bank is able to control the interest rate. But, the interest rate is not the interest rate confronting households and firms; the spread between the two is a critical endogenous variable." However, pre-crisis DSGE models like those in Williamson (1987), Carlstrom and Fuerst (1997), Chari, Christiano, and Eichenbaum (1995), and Christiano, Motto, and Rostagno (2003) and post-crisis DSGE models like Gertler and Karadi (2011), Jermann and Quadrini (2012), Curdia and Woodford (2010), and Christiano, Motto, and Rostagno (2014) offer counterexamples. In all those papers, which are only a subset of the relevant literature, credit and the endogenous spread between the interest rates confronting households and firms play central roles.

Nonlinearities and Lack of Policy Advice

Stiglitz (2017, p. 7) writes that "the large DSGE models that account for some of the more realistic features of the macroeconomy can only be 'solved' for linear approximations and small shocks—precluding the big shocks that take us far away from the domain over which the linear approximation has validity." He (p. 12) writes that "an adequate macro model has to explain how even a moderate shock has large macroeconomic consequences." He claims (p. 1): "[T]he inability of the DSGE model to ... provide policy guidance on how to deal with the consequences [of the crisis], precipitated current dissatisfaction with the model."

Many papers cited throughout this essay offer clear counterexamples to the criticism that dynamic stochastic general equilibrium models don't address nonlinearities and large shocks, or that such models cannot explain why moderate shocks can have large consequences. The claim that DSGE models are unable to provide policy guidance does not square with the simple fact that central banks all over the world actually use DSGE models as part of their policy process.

Heterogeneity

Stiglitz (2017, p. 5) writes that DSGE models do not include heterogeneous agents: "DSGE models seem to take it as a religious tenet that consumption should be explained by a model of a representative agent maximizing his utility over an infinite lifetime without borrowing constraints." This view is obviously at variance with the cutting-edge research in DSGE models discussed earlier.

Dynamic stochastic general equilibrium models will become better as modelers respond to informed criticism. Stiglitz's criticisms are not informed.

Conclusion

The enterprise of dynamic stochastic general equilibrium modeling is an organic process that involves the constant interaction of data and theory. Pre-crisis DSGE models had shortcomings that were highlighted by the financial crisis and its aftermath. Substantial progress has occurred since then. We have emphasized the incorporation of financial frictions and heterogeneity into DSGE models. However, we should also mention that other exciting work is being done in this area, like research on deviations from conventional rational expectations. These deviations include *k*-level thinking, robust control, social learning, adaptive learning, and relaxing the assumption of common knowledge. Frankly, we do not know which of these competing approaches will play a prominent role in the next generation of mainstream DSGE models.

Will the future generation of DSGE models predict the time and nature of the next crisis? Frankly, we doubt it. As far as we know, there is no sure, time-tested way of foreseeing the future. The proximate cause for the financial crisis was a failure across the economics profession, policymakers, regulators, and financial market professionals to recognize and to react appropriately to the growing size and leverage of the shadow-banking sector. DSGE models are evolving in response to that failure as well as to the treasure trove of micro data available to economists. We don't know where that process will lead. But we do know that DSGE models will remain central to how macroeconomists think about aggregate phenomena and policy. There is simply no credible alternative to policy analysis in a world of competing economic forces operating on different parts of the economy.

■ *We are grateful for the comments of Olivier Blanchard, Robert Gordon, Narayana Kocherlakota, Douglas Laxton, Edward Nelson, Giorgio Primiceri, and Sergio Rebelo on an earlier draft of this paper.*

References

Adolfson, Malin, Stefan Laséen, Lawrence J. Christiano, Mathias Trabandt, and Karl Walentin. 2013. "Ramses II—Model Description." Occasional Paper 12, Sveriges Riksbank.

Alpanda, Sami, Gino Cateau, and Césaire Meh. 2014 "A Policy Model to Analyze Macroprudential Regulations and Monetary Policy." Bank of Canada Staff Working Paper 2014-6.

Altig, David, Lawrence J. Christiano, Martin S. Eichenbaum, and Jesper Linde. 2011. "Firm-Specific Capital, Nominal Rigidities and the Business Cycle." *Review of Economic Dynamics* 14(2): 225–47.

Anderson, Eric, Sergio Rebelo, and Arlene Wong. 2018. "Markups across Space and Time." NBER Working Paper 24434.

Angeletos, George-Marios, and Chen Lian. Forthcoming. "Forward Guidance without Common Knowledge." *American Economic Review.*

Argov, Eyal, Emanuel Barnea, Alon Binyamini, Eliezer Borenstein, David Elkayam, and Irit Rozenshtrom. 2012. "MOISE: A DSGE Model for the Israeli Economy." Research Discussion Paper 2012.06, Bank of Israel, April 4.

Auclert, Adrien. 2017. "Monetary Policy and the Redistribution Channel." NBER Working Paper 23451.

Backus, David K., Patrick J. Kehoe, and Finn E. Kydland. 1992. "International Real Business Cycles." *Journal of Political Economy* 100(4): 745–75.

Benes, Jaromir, Michael Kumhof, and Douglas Laxton. 2014. "Financial Crises in DSGE Models: A Prototype Model." International Monetary Fund Working Paper 14/57.

Benhabib, Jess, Stephanie Schmitt-Grohé, and Martín Uribe. 2001. "Monetary Policy and Multiple Equilibria." *American Economic Review* 91(1): 167–86.

Berger, David, Veronica Guerrieri, Guido Lorenzoni, and Joseph Vavra. Forthcoming. "House Prices and Consumer Spending." *Review of Economic Studies.*

Bernanke, Ben S. 1995. "The Macroeconomics of the Great Depression: A Comparative

Approach." *Journal of Money, Credit and Banking* 27(1): 1–28.

Bernanke, Ben S. 2009. "Opening Remarks: Reflections on a Year of Crisis." Federal Reserve Bank of Kansas City's Economic Policy Symposium, Jackson Hole, Wyoming, August 21, 2009.

Bernanke, Ben S. 2010. "Statement Before the Financial Crisis Inquiry Commission, Washington, D.C." September 2. http://www.federalreserve.gov/newsevents/testimony/bernanke20100902a.pdf.

Bernanke, Ben S., and Alan S. Blinder. 1992. "The Federal Funds Rate and the Channels of Monetary Transmission." *American Economic Review* 82(4): 901–21.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." Chap. 21 in *Handbook of Macroeconomics* vol. 1C. Elsevier.

Bianchi, Javier, Juan Carlos Hatchondo, and Leonardo Martinez. 2016. "International Reserves and Rollover Risk." Working Paper 735, Federal Reserve Bank of Minneapolis.

Blanchard, Olivier, Christopher J. Erceg, and Jesper Lindé. 2017. "Jump-Starting the Euro-Area Recovery: Would a Rise in Core Fiscal Spending Help the Periphery?" *NBER Macroeconomics Annual* 2017, vol. 31, pp. 103–182.

Boldrin, Michele, Lawrence J. Christiano, and Jonas D. M. Fisher. 2001. "Habit Persistence, Asset Returns, and the Business Cycle." *American Economic Review* 91(1): 149–66.

Boneva, Lena Mareen, R. Anton Braun, and Yuichiro Waki. 2016. "Some U npleasant Properties of Loglinearized Solutions when the Nominal Rate is Zero." *Journal of Monetary Economics* 84: 216–232.

Brzoza-Brzezina, Michał, and Marcin Kolasa. 2013. "Bayesian Evaluation of DSGE Models with Financial Frictions." *Journal of Money, Credit and Banking* 45(8): 1451–76.

Caballero, Ricardo J., and Arvind Krishnamurthy. 2008. "Collective Risk Management in a Flight to Quality Episode." *Journal of Finance* 63(5): 2195–2230.

Cai, Michael, Marco Del Negro, Marc P. Giannoni, Abhi Gupta, Pearl Li, and Erica Moszkowski. 2018. "DSGE Forecasts of the Lost Recovery." Federal Reserve Bank of New York Staff Reports 844.

Calvo, Guillermo A. 1983. "Staggered Prices in a Utility-Maximizing Framework." *Journal of Monetary Economics* 12(3): 383–98.

Canova, Fabio, and Luca Sala. 2009. "Back to Square One: Identification Issues in DSGE Models." *Journal of Monetary Economics* 56(4): 431–49.

Carabenciov, Ioan, Charles Freedman, Roberto Garcia-Saltos, Douglas Laxton, Ondra Kamenik, and Petar Manchev. 2013. "GPM6: The Global Projection Model with 6 Regions." IMF Working Paper 13/87.

Carlstrom, Charles T., and Timothy S. Fuerst. 1997. "Agency Costs, Net Worth, and Business Fluctuations: A Computable General Equilibrium Analysis." *American Economic Review* 87(5): 893–910.

Carlstrom, Charles T., Timothy S. Fuerst, and Matthias Paustian. 2015. "Inflation and Output in New Keynesian Models with a Transient Interest Rate Peg." *Journal of Monetary Economics* 76: 230–43.

Carroll, Christopher D., Jody Overland, and David N. Weil. 1997. "Comparison Utility in a Growth Model." *Journal of Economic Growth* 2(4): 339–67.

Carroll, Christopher D., Jody Overland, and David N. Weil. 2000. "Saving and Growth with Habit Formation." *American Economic Review* 90(3): 341–55.

Chari, V. V., Lawrence J. Christiano, and Martin S. Eichenbaum. 1995. "Inside Money, Outside Money, and Short-term Interest Rates." *Journal of Money, Credit and Banking* 27(4): 1354–1402.

Chernozhukov, Victor, and Han Hong. 2003. "An MCMC Approach to Classical Estimation." *Journal of Econometrics* 115(2): 293–346.

Chetty, Raj, Adam Guren, Day Manoli, and Andrea Weber. 2011. "Are Micro and Macro Labor Supply Elasticities Consistent? A Review of Evidence on the Intensive and Extensive Margins." *American Economic Review* 101(3): 471–75.

Christiano, Lawrence J., and Martin S. Eichenbaum. 1992. "Current Real-Business-Cycle Theories and Aggregate Labor Market Fluctuations." *American Economic Review* 82(3): 430–50.

Christiano, Lawrence J., and Martin S. Eichenbaum. 2012. "Notes on Linear Approximations, Equilibrium Multiplicity and E-Learnability in the Analysis of the Zero Lower Bound." http://faculty.wcas.northwestern.edu/~lchrist/research/Zero_Bound/webpage.html.

Christiano, Lawrence J., Martin S. Eichenbaum, and Charles L. Evans. 1996. "The Effects of Monetary Policy Shocks: Evidence from the Flow of Funds." *Review of Economics and Statistics* 78(1): 16–34.

Christiano, Lawrence J., Martin S. Eichenbaum, and Charles L. Evans. 1999. "Monetary Policy Shocks: What Have We Learned and to What End?" Chap. 2 in *Handbook of Macroeconomics* vol. 1A, edited by John B. Taylor and Michael Woodford. North-Holland, Elsevier.

Christiano, Lawrence J., Martin S. Eichenbaum, and Charles L. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *Journal of Political Economy* 113(1): 1–45.

Christiano, Lawrence J., Martin S. Eichenbaum, and Benjamin K. Johannsen. 2018. "Does the New Keynesian Model Have a Uniqueness Problem?" NBER Working Paper 24612.

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

Christiano, Lawrence J., Martin S. Eichenbaum, and Mathias Trabandt. 2015. "Understanding the Great Recession." *American Economic Journal: Macroeconomics* 7(1): 110–67.

Christiano, Lawrence J., Martin S. Eichenbaum, and Mathias Trabandt. 2016. "Unemployment and Business Cycles." *Econometrica* 84(4): 1523–69.

Christiano, Lawrence J., Roberto Motto, and Massimo Rostagno. 2003. "The Great Depression and the Friedman-Schwartz Hypothesis." *Journal of Money, Banking, and Credit* 25(6).

Christiano, Lawrence J., Roberto Motto, and Massimo Rostagno. 2014. "Risk Shocks." *American Economic Review* 104(1): 27–65.

Christiano, Lawrence J., Mathias Trabandt, and Karl Walentin. 2010. "DSGE Models for Monetary Policy Analysis." Chap. 7 in *Handbook of Monetary Economics*, vol. 3, edited by Benjamin M. Friedman and Michael Woodford. Elsevier.

Clarida, Richard, Jordi Galí, and Mark Gertler. 1999. "The Science of Monetary Policy: A New Keynesian Perspective." *Journal of Economic Literature* 37(2): 1661–1707.

Clinton, Kevin, Tibor Hlédik, Tomáš Holub, Douglas Laxton, and Hou Wang. 2017. "Czech Magic: Implementing Inflation-Forecast Targeting at the CNB." International Monetary Fund Working Paper no. 17/21.

Coenen, Günter et al. 2012. "Effects of Fiscal Stimulus in Structural Models." *American Economic Journal: Macroeconomics* 4(1): 22–68.

Cole, Harold L., and Timothy J. Kehoe. 2000. "Self-Fulfilling Debt Crises." *Review of Economic Studies* 67(1): 91–116.

Comin, Diego, and Mark Gertler. 2006. "Medium-Term Business Cycles." *American Economic Review* 96(3): 523–51.

Constantinides, George M. 1990. "Habit Formation: A Resolution of the Equity Premium Puzzle." *Journal of Political Economy* 98(3): 519–43.

Cúrdia, Vasco, and Michael Woodford. 2010. "Credit Spreads and Monetary Policy." *Journal of Money, Credit and Banking* 42(s1): 3–35.

Del Negro, Marco, Marc Giannoni, and Christina Patterson. 2012. "The Forward Guidance Puzzle." Federal Reserve Bank of New York Staff Report 574.

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

Dorich, José, Michael K. Johnston, Rhys R. Mendes, Stephen Murchison, and Yang Zhang. 2013. "ToTEM II: An Updated Version of the Bank of Canada's Quarterly Projection Model." Bank of Canada Technical Report 100.

Eberly, Janice, Sergio Rebelo, and Nicolas Vincent. 2012. "What Explains the Lagged-Investment Effect?" *Journal of Monetary Economics* 59(4): 370–80.

Eggertsson, Gauti B., and Michael Woodford. 2003. "The Zero Bound on Interest Rates and Optimal Monetary Policy." *Brookings Papers on Economic Activity*, no. 1.

Eichenbaum, Martin S., and Lars Peter Hansen. 1990. "Estimating Models with Intertemporal Substitution Using Aggregate Time Series Data." *Journal of Business & Economic Statistics* 8(1): 53–69.

Eichenbaum, Martin, Nir Jaimovich, and Sergio Rebelo. 2011. "Reference Prices, Costs, and Nominal Rigidities." *American Economic Review* 101(1): 234–62.

Erceg, Christopher J., Luca Guerrieri, and Christopher Gust. 2006. "SIGMA: A New Open Economy Model for Policy Analysis." *International Journal of Central Banking* 2(1).

Erceg, Christopher, and Jesper Lindé. 2014. "Is There a Fiscal Free Lunch in a Liquidity Trap?" *Journal of the European Economic Association* 12(1): 73–107.

Farhi, Emmanuel, and Iván Werning. 2017. "Monetary Policy, Bounded Rationality, and Incomplete Markets." NBER Working Paper 23281.

Fernández-Villaverde, Jesús, Juan Francisco Rubio-Ramirez, and Frank Schorfheide. 2016. "Solution and Estimation Methods for DSGE Models." Chap. 10 in *Handbook of Macroeconomics* 2A, edited by John B. Taylor and Harald Uhlig. North-Holland, Elsevier.

Fischer, Stanley. 2017. "Stanley Fisher: I'd Rather Have Bob Solow than an Econometric Model, But ..." Speech at the Warwick Economics Summit, Coventry, United Kingdom, February, 11. http://www.bis.org/review/r170214a.htm.

Freedman, Charles, Michael Kumhof, Douglas Laxton, Dirk Muir, and Susanna Mursula. 2009. "Fiscal Stimulus to the Rescue? Short-Run Benefits and Potential Long-Run Costs of Fiscal Deficits." IMF Working Paper 09/255.

Friedman, Milton. 1959. U.S. Congress, Joint Economic Committee, Hearings, Employment, Growth, and Price Levels, Part 4 (86th Cong., 1st sess., 1959), pp. 615–16.

Friedman, Milton. 1968. "The Role of Monetary Policy." *American Economic Review* 58(1): 1–17.

Friedman, Milton, and Anna Jacobson Schwartz. 1963. *A Monetary History of the United*

States, 1867–1960. Princeton University Press.

Fuhrer, Jeffrey, C. 2000. “Habit Formation in Consumption and Its Implications for Monetary Policy Models.” *American Economic Review* 90(3): 367–90.

Gabaix, Xavier. 2016. “A Behavioral New Keynesian Model.” NBER Working Paper 22954.

Galí, Jordi, J. David López-Salido, and Javier Vallés. 2007. “Understanding the Effects of Government Spending on Consumption.” *Journal of the European Economic Association* 5(1): 227–70.

Gerdrup, Karsten R., Erling Motzfeldt Kravik, Kenneth Sæterhagen Paulsen, and Ørjan Robstad. 2017. “Documentation of NEMO—Norges Bank’s Core Model for Monetary Policy Analysis and Forecasting.” Staff Memo no. 8, Norges Bank.

Gertler, Mark, and Peter Karadi. 2011. “A Model of Unconventional Monetary Policy.” *Journal of Monetary Economics* 58(1): 17–34.

Gertler, Mark, and Peter Karadi. 2015. “Monetary Policy Surprises, Credit Costs, and Economic Activity.” *American Economic Journal: Macroeconomics* 7(1): 44–76.

Gertler, Mark, and Nobuhiro Kiyotaki. 2015. “Banking, Liquidity, and Bank Runs in an Infinite Horizon Economy.” *American Economic Review* 105(7): 2011–2043.

Gertler, Mark , Nobuhiro Kiyotaki, and Andrea Prestipino. 2016. “Wholesale Banking and Bank Runs in Macroeconomic Modeling of Financial Crises.” Chap. 16 in *Handbook of Macroeconomics,* vol. 2B, edited by John Taylor and Harald Uhlig. Elsevier.

Gilchrist, Simon, Raphael Schoenle, Jae Sim, and Egon Zakrajšek. 2017. “Inflation Dynamics during the Financial Crisis.” *American Economic Review* 107(3): 785–823.

Gornemann, Nils, Keith Kuester, and Makoto Nakajima. 2016. “Doves for the Rich, Hawks for the Poor? Distributional Consequences of Monetary Policy.” International Finance Discussion Papers 1167.

Gremel, Christina M., Jessica H. Chancey, Brady K. Atwood, Guoxiang Luo, Rachael Neve, Charu Ramakrishnan, Karl Deisseroth, David M. Lovinger, and Rui M. Costa. 2016. “Endocannabinoid Modulation of Orbitostriatal Circuits Gates Habit Formation.” *Neuron* 90(6): 1312–24.

Guerrieri, Veronica, and Guido Lorenzoni. 2017. “Credit Crises, Precautionary Savings, and the Liquidity Trap.” *Quarterly Journal of Economics* 132(3): 1427–67.

Gürkaynak, Refet S., Brian Sack, and Eric T. Swanson. 2005. “Do Actions Speak Louder than Words? The Response of Asset Prices to Monetary Policy Actions and Statements.” *International Journal of Central Banking* 1(1): 55–93.

Gust, Christopher, Edward Herbst, David López-Salido, and Matthew E. Smith. 2017. “The Empirical Implications of the Interest-Rate Lower Bound.” *American Economic Review* 107(7): 1971–2006.

Hansen, Lars Peter. 1982. “Large Sample Properties of Generalized Method of Moments Estimators.” *Econometrica* 50(4): 1029–54.

Hume, David. 1742 [1987]. “Of Money,” Part II, Essay III.7, in *Essays, Moral, Political, and Literary,* edited by Eugene F. Miller. Liberty Fund. http://www.econlib.org/library/LFBooks/Hume/hmMPL26.html.

Iacoviello, Matteo. 2005. “House Prices, Borrowing Constraints, and Monetary Policy in the Business Cycle.” *American Economic Review* 95(3): 739–64.

Iacoviello, Matteo, and Stefano Neri. 2010. “Housing Market Spillovers: Evidence from an Estimated DSGE Model.” *American Economic Journal: Macroeconomics* 2(2): 125–64.

Jermann, Urban, and Vincenzo Quadrini. 2012. “Macroeconomic Effects of Financial Shocks.” *American Economic Review* 102(1): 238–71.

Justiniano, Alejandro, Giorgio E. Primiceri, and Andrea Tambalotti. 2010. “Investment Shocks and Business Cycles.” *Journal of Monetary Economics* 57(2): 132–45.

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante. 2018. “Monetary Policy according to HANK.” *American Economic Review* 108(3): 697–743.

Kiley, Michael. 2016. “Policy Paradoxes in the New-Keynesian Model.” *Review of Economic Dynamics* 21: 1–15.

King, Robert G., and Sergio T. Rebelo. 1999. “Resuscitating Real Business Cycles.” Chap. 14 in *Handbook of Macroeconomics,* vol. 1, pp. 927–1007. Elsevier.

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

Kocherlakota, Narayana R. 2000. “Creating Business Cycles through Credit Constraints.” *Quarterly Review,* Federal Reserve Bank of Minneapolis 24(3): 2–10.

Korinek, Anton. 2017. “Thoughts on DSGE Macroeconomics: Matching the Moment, But Missing the Point?” Available at SSRN: https://ssrn.com/abstract=3022009.

Krugman, Paul R. 1998. “It’s Baaack: Japan’s Slump and the Return of the Liquidity Trap.” *Brookings Papers on Economic Activity* no. 2, pp. 137–205.

Kuttner, Kenneth. 2001. “Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market.” *Journal of Monetary Economics* 47(3): 523–44.

Kydland, Finn E., and Edward C. Prescott. 1982. “Time to Build and Aggregate Fluctuations.”

Econometrica 50(6): 1345–70.

Lindé, Jesper, Frank Smets, and Rafael Wouters. 2016. "Challenges for Central Bank Models." *Handbook of Macroeconomics,* vol. 2B, pp. 2185–2256. North-Holland.

Lindé, Jesper, and Mathias Trabandt. 2018. "Resolving the Missing Deflation Puzzle." https://sites.google.com/site/mathiastrabandt/home/downloads/LindeTrabandt_Inflation.pdf.

Lindé, Jesper, and Mathias Trabandt. Forthcoming**.** "Should We Use Linearized Models to Calculate Fiscal Multipliers?" *Journal of Applied Econometrics.*

Liu, Zheng, Pengfei Wang, and Tao Zha. 2013. "Land-Price Dynamics and Macroeconomic Fluctuations." *Econometrica* 81(3): 1147–84.

Long, John B., and Charles I. Plosser. 1983. "Real Business Cycles." *Journal of Political Economy* 91(1): 39–69.

Lucas, Robert E., Jr., and Edward C. Prescott. 1971. "Investment under Uncertainty." *Econometrica* 39(5): 659–81.

Lucca, David Olivier. 2006. "Essays in Investment and Macroeconomics." PhD Dissertation, Northwestern University, Department of Economics.

Matsuyama, Kiminori. 1984. "A Learning Effect Model of Investment: An Alternative Interpretation of Tobin's Q." Unpublished paper, Northwestern University.

McKay, Alisdair, Emi Nakamura, and Jón Steinsson. 2016. "The Power of Forward Guidance Revisited." *American Economic Review* 106(10): 3133–58.

McKay, Alisdair, and Ricardo Reis. 2016. "The Role of Automatic Stabilizers in the U.S. Business Cycle." *Econometrica* 84(1): 141–94.

Mertens, Karel, and Morten O. Ravn. 2013. "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States." *American Economic Review* 103(4): 1212–47.

Mussa, Michael. 1986. "Nominal Exchange Rate Regimes and the Behavior of Real Exchange Rates: Evidence and Implications." *Carnegie-Rochester Conference Series on Public Policy* 25(1): 117–214.

Nakamura, Emi, and Jón Steinsson. 2008. "Five Facts about Prices: A Reevaluation of Menu Cost Models." *Quarterly Journal of Economics* 123(4): 1415–64.

Oh, Hyunseung, and Ricardo Reis. 2012. "Targeted Transfers and the Fiscal Response to the Great Recession." *Journal of Monetary Economics* 59(Supplement): S50–S64.

Ottonello, Pablo, and Thomas Winberry. 2018. "Financial Heterogeneity and the Investment Channel of Monetary Policy." NBER Working Paper 24221.

Rotemberg, Julio J., and Michael Woodford. 1991. "Markups and the Business Cycle." *NBER Macroeconomics Annual* 6: 63–140.

Rudolf, Barbara, and Mathias Zurlinden. 2014. "A Compact Open Economy DSGE Model for Switzerland." SNB Economic Studies, no. 2014-08, Swiss National Bank.

Sims, Christopher A. 1986. "Are Forecasting Models Usable for Policy Analysis?" *Federal Reserve Bank of Minneapolis Quarterly Review* 10(1): 2–16.

Smets, Frank, Kai Christoffel, Günter Coenen, Roberto Motto, and Massimo Rostagno. 2010. "DSGE Models and Their Use at the ECB." *SERIEs* 1(1–2): 51–65.

Smets, Frank, and Raf Wouters. 2003. "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area." *Journal of the European Economic Association* 1(5): 1123–75.

Smets, Frank, and Rafael Wouters. 2007. "Shocks and Frictions in US Business Cycles: A Baysian DSGE Approach." *American Economic Review* 97(3): 586–606.

Stiglitz, Joseph E. 2017. "Where Modern Macroeconomics Went Wrong." NBER Working Paper 23795.

Taylor, John B. and Harald Uhlig, eds. 1990. "Solving Nonlinear Rational Expectations Models." A set of papers in *Journal of Business and Economic Statistics* 8(1).

Williamson, Stephen D. 1987. "Financial Intermediation, Business Failures, and Real Business Cycles." *Journal of Political Economy* 95(6): 1196–1216.

Woodford, Michael. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy.* Princeton University Press.

Woodford, Michael. 2012. "Methods of Policy Accommodation at the Interest-rate Lower Bound." *Proceedings—Economic Policy Symposium—Jackson Hole, Federal Reserve Bank of Kansas City*, pp. 185–288.

Woodford, Michael. 2018. "Monetary Policy Analysis when Planning Horizons Are Finite." *NBER Macro Annual 2018,* vol. 33, edited by Martin Eichenbaum and Jonathan A. Parker. University of Chicago Press.

Yun, Tack. 1996. "Nominal Price Rigidity, Money Supply Endogeneity, and Business Cycles." *Journal of Monetary Economics* 37(2): 345–70.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 141–166

Evolution of Modern Business Cycle Models: Accounting for the Great Recession

Patrick J. Kehoe, Virgiliu Midrigan, and Elena Pastorino

Modern business cycle theory focuses on the study of dynamic stochastic general equilibrium (DSGE) models that generate aggregate fluctuations similar to those experienced by actual economies. We discuss how these modern business cycle models have evolved across three generations, from their roots in the early real business cycle models of the late 1970s through the turmoil of the Great Recession four decades later.

The first generation models were real (that is, without a monetary sector) business cycle models that primarily explored whether a small number of shocks, often one or two, could generate fluctuations similar to those observed in aggregate variables such as output, consumption, investment, and hours. These basic models disciplined their key parameters with micro evidence and were remarkably successful in matching these aggregate variables.

Over time, as the theory evolved and computational possibilities expanded, a second generation of these models appeared. These models incorporated frictions such as sticky prices and wages, and were primarily developed to be used in central banks for short-term forecasting purposes and for performing counterfactual policy

■ *Patrick Kehoe is a Professor of Economics at Stanford University, Stanford, California; Professor of Economics, University College London, London, United Kingdom; and a consultant at the Federal Reserve Bank of Minneapolis, Minneapolis, Minnesota. Virgiliu Midrigan is a Professor of Economics at New York University, New York, New York. Elena Pastorino is a Research Fellow at the Hoover Institution, Stanford University, a Faculty Research Fellow at the Economics Department, Stanford University, and a Research Fellow at the Stanford Institute for Economic Policy Research, all in Stanford, California.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.141 doi=10.1257/jep.32.3.141

experiments. Due to the focus on forecasting, the main emphasis in choosing the parameters of these models was on their ability to match the behavior of, say, 10 to 12 aggregate variables rather than on carefully matching them up with micro evidence. Although these models were called New Keynesian, they had little to do with traditional Keynesian models. Rather, they were simply real business cycle models augmented with sticky prices and wages. Indeed, a canonical real business cycle model augmented with money and flexible prices—so that monetary policy can be meaningfully discussed—has essentially the same implications for the importance of various shocks for business cycles, and nearly identical implications for optimal monetary and fiscal policy, as these New Keynesian models do.

During the last decade or so, macroeconomists working on the next generation of business cycle models have benefited from the development of new algorithms and the increase in computing power to incorporate the rich heterogeneity of patterns from the micro data, even in models where no aggregation to a representative firm or consumer is feasible. A defining characteristic of these models is not the heterogeneity among model agents they accommodate nor the micro-level evidence they rely on (although both are common), but rather the insistence that any new parameters or feature included be explicitly disciplined by direct evidence. The spirit of the discipline of the parameters of these third-generation models echoes the attitudes of the original developers of first-generation models, except that third-generation models are sophisticated enough to match a wealth of additional aspects of the micro data. The growth of such third-generation models was hastened by the Great Recession, a striking episode that led macroeconomists to dig more deeply than before into the links between disruptions in the process of financial intermediation and business cycles.

We briefly review the development of business cycle models through its three generations. We then show how two versions of this latest generation of modern business cycle models, which are real business cycle models with frictions in labor and financial markets, can account, respectively, for the aggregate and the cross-regional fluctuations observed in the United States during the Great Recession. We begin with a comparison of the comovements of the major macro aggregates in the two largest postwar recessions in the United States: the 1982 recession, which exhibited essentially no financial distress, and the Great Recession, which was characterized by the greatest distress of any post–World War II business cycle. In the 1982 recession, the drop in measured total factor productivity was as large as the drop in output, whereas hours fell much less than output—a pattern that holds up across most postwar recessions. The pattern of these variables in the Great Recession was strikingly different: measured total factor productivity barely fell, whereas labor fell more than output. We argue that this fundamental difference in the pattern of comovements of output, measured productivity, and hours in the Great Recession, along with the documented increase in financial distress, calls for new mechanisms to explain this downturn.

Along with these specific examples, our overall message is that the basic questions that macroeconomists address have not changed over time, but rather that the development of real business cycle methods has fundamentally changed how these questions are posed and answered. Now, we no longer ask, “What is the best policy action we can take today?” but instead ask, “How does the behavior of the economy considered compare under one rule for policy versus another rule for policy?” We answer these questions with models that are specified at a deep enough level that their primitive features are plausibly invariant to the policy exercises of interest and are consistent with a wealth of relevant micro evidence.

Of course, macroeconomists still hold widely divergent views about the answers to fundamental questions such as the ability of a well-specified rule for monetary policy to singlehandedly stop an incipient Great Depression episode in its tracks. But we now agree that a disciplined debate of such questions rests on communication in the language of a dynamic general equilibrium model, specified at the level of primitives, that is internally as well as externally validated. Through such a disciplined communication, we can reduce endless debates about opinions to concrete ones about the evidence for detailed mechanisms and parameters.

First Generation: Basic Real Business Cycle Models

Modern business cycle models were developed in response to the “Lucas critique” of large-scale econometric models built along traditional Keynesian lines, which were the dominant scientific paradigm in macroeconomics from the 1950s to the 1970s. Lucas (1976) argued that unless an econometric model is built at a deep enough level so that its parameters are invariant to the class of policy interventions being considered, the model is of no value in assessing policy interventions, regardless of how well the model performs in unconditional forecasting. The reason is that the policy intervention may affect key parameters that are presumed to be constant, and so invalidate the policy exercise. This critique motivated macroeconomists to micro-found their dynamic models by building them starting from the specification of technology, preferences, and other primitive constraints, along with an equilibrium concept. In the resulting equilibrium, agents think intertemporally, not just statically, and decisions on consumption, investment, and labor supply must simultaneously satisfy resource constraints and budget constraints.

From a practical viewpoint, the Keynesian macroeconometric framework fell out of favor after the period of stagflation that many developed economies experienced during the 1970s when it became clear that this framework offered neither a cohesive theoretical explanation for this stagflation episode nor, in light of the Lucas critique, defensible policy advice.

Early Attempts to Match Aggregate Macro Variables

The first generation of modern business cycle models consisted of simple, primarily real business cycle models with a representative consumer and few frictions. These models were used to examine the extent to which a small number of shocks, say, one or two, could account for the movements of major aggregates, such as output, consumption, investment, and hours. In this framework, all economic behavior was derived from one general equilibrium model in which agents adjust their behavior when policies, specified as rules, are changed and forecast the future using the true probability distributions implied by the model. A prominent example of such a model is in Prescott (1986), which features a representative consumer, one real exogenous technology shock, frictionless markets, and no money or nominal variables. Earlier versions of first-generation business cycle models by Kydland and Prescott (1982) and Long and Plosser (1983) featured much more complex real sides of the economy.

The early papers in this vein documented patterns in the macro data, typically summarized by a table of moments of the data such as standard deviations, autocorrelations, and cross-correlations of output, consumption, investment, hours, and a few other variables. This table, often referred to as a "KP table" in light of Kydland and Prescott (1982), compared these moments in the data to those generated by the model. Key discrepancies between the predictions of the model and the actual data were often referred to as an "anomaly." For example, a key anomaly of the early work was that hours in the models fluctuated by less than half as much as in the data.

A typical research paper of this generation usefully focused on showing how adding one new mechanism to an otherwise standard model helped the model to better account for some existing anomaly. Over time, the work evolved into the study of richer mechanisms. The work in this vein did make a serious attempt to discipline new parameters with external evidence, which often involved connecting the new parameters considered to those from micro studies. Nonetheless, the models were sufficiently simple that it was often difficult to connect tightly the features or implications of new mechanisms with the requisite evidence from micro-level data on consumers and firms.

The basic real business cycle models of that time, with only stochastic movements in total factor productivity, generated fluctuations in the major aggregates largely in accord with those observed in the data when the technology parameter was simply taken as exogenous. As the Kydland and Prescott (1982) paper makes abundantly clear, they were surprised that their simple model, which abstracted from monetary frictions, could do so. However, this approach also gave rise to some oft-repeated questions about their purpose and design, which we discuss next.

Why Abstract from Money?

Why did early practitioners of the real business cycle approach focus on models that abstracted from money and, hence, monetary policy? There are three reasons. First, the goal of this early work was to develop the simplest possible model based on a coherent set of assumptions, in which agents acted in their own self-interest and that could produce fluctuations in aggregate quantities, such as output, consumption, investment, and hours, broadly in accordance with those in the data. Second, the simplicity of the models reflected existing limits on the ability to compute these models numerically. In the late 1970s, macroeconomists lacked the methods and computing power to solve complicated models with heterogeneous agents, multiple frictions, and nonlinear effects. Third, many of the macroeconomists working on real business cycles were deeply affected by the failures of the policy advice derived from the earlier statistical Keynesian models that helped to exacerbate the stagflation of the 1970s. Hence, they retreated to a humbler intellectual position focused on building coherent foundations for macroeconomics and avoiding both the Lucas critique and the hubris that led to previous mistakes. Most macroeconomists felt quite uncomfortable rushing back to the policy arena without well-developed models.

Sometimes it is thought that the most important reason why real business cycle modelers abstracted from money was because they believed that monetary policy has no effect on the real economy (for example, see Summers 1986; Romer 2016). We argue that this view is incorrect. Part of the misunderstanding seems to have arisen from the well-known policy ineffectiveness proposition of Sargent and Wallace (1975). Sargent and Wallace articulated a critique of models that produced real effects of money solely because agents were assumed to be irrational.

In a similar vein, Barro (1977) critiques "sticky wage" models. He argues that even if nominal wages do not vary with monetary shocks, if we model wages and employment levels as part of a contract that is agreed upon mutually by firms and workers, then there is no room for monetary feedback rules to systematically improve outcomes. Barro points out a weak theoretical link in a popular mechanism: even if sticky nominal wages are assumed, existing models generate real effects solely because they assume that an employment contract does not specify hours worked in addition to wages and so leaves unexploited mutual gains from trade.

Properly understood, both the Sargent and Wallace (1975) and Barro (1977) papers were critiques of popular existing mechanisms for monetary nonneutrality, rather than the expression of either a belief that no coherent model could be developed in which monetary policy had real effects or that in actual economies monetary policies have no real effects. For example, there is near-universal agreement that the disastrous monetary policies pursued by countries such as Argentina, Brazil, and Chile had serious adverse effects on these economies.

More broadly, macroeconomists agree on the direction in which monetary policy should respond to shocks over the cycle—although they disagree on the magnitude of desirable monetary interventions. Even with a general agreement on how monetary policy works in a qualitative sense, it remains an exceptionally difficult task to build a coherent model in which consumers and firms act in their own self-interest that quantitatively captures well how monetary policy works. For instance, as Barro (1977) foresaw, the difficulty with many of the sticky wage or price models is that they rely on agents agreeing to contracts that ignore mutual gains for trade. At a deeper level, the fact that a contract in such models is not optimal given preferences, technology, and information makes them subject to the Lucas critique. Our own sense is that depending on the exact standard that needs to be met before the word "coherent" is applicable, macroeconomists may still be far away from achieving the goal of a coherent model.

Why Focus on Technology Shocks?

Real business cycle models are driven by what are commonly referred to as technology shocks. Why did early researchers choose to treat the aggregate productivity parameter in the output production function as the key stochastic variable? There are two practical reasons. First, productivity is relatively easy to measure given a functional form assumption for the aggregate production function and data on aggregate output, the capital stock, and hours. Second, with a single shock added nearly anywhere else in a one-sector growth model, it is difficult to generate the business cycle comovements between output, consumption, investment, and hours found in the data.

For example, a shock that primarily leads to a deep fall in investment tends to make consumption and investment move in opposite directions, which is inconsistent with the data. To see why, consider the effect of a fall in investment on output. Since the capital stock is over ten times investment, a drop in investment has only a tiny effect on the capital stock and no direct effect on labor, so the amount produced with capital and labor barely moves. But from the resource constraint, consumption and investment must add up to output. Hence, the only way that investment can fall a lot, output barely move, and the resource constraint be satisfied is for consumption to rise. Using a quantitative model, Cooper and Ejarque (2000) show that shocks that operate through an investment channel counterfactually imply that consumption and investment are negatively correlated.

Consider next a shock that reduces the desire to work and, hence, reduces hours worked. With a Cobb–Douglas aggregate production technology and a labor share of two-thirds, a given percentage drop in hours, say 10 percent, leads to only two-thirds (6.7 percent) as large a drop in output. But if such shocks are the main drivers of the business cycle, then labor would be much more volatile than output, an implication that is inconsistent with US business cycles prior to the Great Recession (Chari, Kehoe, and McGrattan 2007; Brinca, Chari, Kehoe, and McGrattan 2016).

Finally, it is also important to understand how this time-varying aggregate productivity parameter should be interpreted. From the beginning of real business cycle theory, it was well accepted that movements in this parameter should not be interpreted as changes in "technology." That is, falls in measured total factor productivity should not be thought of as individual firms forgetting how to produce or deteriorations in the blueprints at the firm level for turning capital and labor into output. Rather, the time variation of the productivity parameter has always been thought of as a stand-in for deeper models of how economic outputs and inputs adjust to various nonproductivity shocks.

For one example, Lagos (2006) shows that in a standard search model with only firm-level productivity differences, an increase in either employment subsidies or costs of firing workers decreases the cutoff for how productive an individual firm must be in order to operate. Hence, these policies lead the average productivity of firms to fall and, hence, lead to a fall in total factor productivity. For another example, Chari, Kehoe, and McGrattan (2007) consider an increase in input financing frictions across sectors such that in bad times the cost of borrowing in some sectors rises relative to that in other sectors. This financing friction distorts the mix of each sector's inputs in final output and hence gives rise to measured falls in total factor productivity.

An alternative view is that neither of these approaches is necessary to understand drops in measured total factor productivity because this measured drop mostly disappears if we simply adjust for the fall in capital utilization, u_{Kt}, in downturns. (To see how this argument works, let $u_{Kt}K_t$ be the service flow from the capital stock, K_t, and $Y_t = A_t \left(u_{Kt}\ K_t\right)^a L_t^{1-\alpha}$ be a Cobb–Douglas production function. Clearly, drops in u_{Kt}^{α} show up as drops in total factor productivity.)

The challenge for this view is to provide a micro-founded reason for utilization to fall steeply enough during recessions to account for the measured fall in total factor productivity. The problem is that given that the vast bulk of a firm's expenses is for labor, keeping the capital stock running is typically much less expensive than paying for labor. Hence, quantitatively relevant micro-founded models often imply very modest declines in capital utilization during downturns. Moreover, since the capital share is small, say, α = 25 percent, such falls in capital utilization can account for only a very small fraction of the measured fall in total factor productivity. Of course, if in the data, firms actually drastically reduce their capital utilization in recessions, then the puzzle is to explain why they do so. More theoretical work needs to be done in this area for progress to be made.

Second Generation: Real Business Cycle Models for Central Banks with a New Keynesian Twist

The *second generation* of modern business cycle models consisted of medium-scale dynamic stochastic general equilibrium models, which were nearly all of the New Keynesian variety and much more complex than those of the first generation. The development of these models was driven by a desire from central banks around the world to find a replacement for the discredited large-scale Keynesian models. As a result, this new generation of medium-scale New Keynesian models needed to be conceived in a way that money could have real effects and be sophisticated enough that they could be used for forecasting.

These second-generation models were designed to fit the behavior of 10 or so aggregate time series that include output, consumption, investment, hours, and some nominal variables such as inflation and nominal interest rates. Because the metric for success of these models was their ability to reproduce the behavior of these aggregates, most of the effort in these models was expended on adding additional features—such as one shock per equation, nonstandard adjustment costs, and extra parameters in preferences and technology—that allowed the model to fit in-sample properties of these aggregates. Little effort was devoted to ensuring that the added shocks, especially the unobservable ones, were clearly interpretable and that the added parameters were disciplined by an explicit attempt to validate them. In practice, these models featured such a complex mix of competing mechanisms, frictions, and shocks that they were quite difficult to understand. In this sense, the methodology for building and assessing second-generation modern business cycle models diverged sharply from that of first-generation models.

A more fundamental methodological issue with these second-generation models, which even now deeply divides macroeconomists, is how to build a model that is not subject to the Lucas critique. In practice, this means we need to ask, "What is a primitive enough level at which to specify a model so that the resulting model is arguably invariant to the policy interventions of interest?" For first-generation modelers, this level consists of technologies, including commitment technologies, preferences, information structure, and physical constraints, such as capital adjustment costs. After these objects are specified, agents are free to choose the contracts to sign or the assets to trade, subject to these primitive constraints. Second-generation modelers, instead, appended direct restrictions on contracts, such as particular forms of sticky wage contracts or restrictions on the class of asset trades allowed, even though these restrictions are not in any agent's interest given the primitive constraints.

For example, a second-generation modern business cycle model might assume that private contracts cannot depend on observable variables outside of any single agent's control, such as aggregate output, and then argue that such a restriction justifies government intervention in the form of a state-contingent tax policy to partially restore the effective insurance not provided by private contracts. From the point of view of a first-generation modeler, this approach is problematic, since the government intervention may affect the unspecified premise of why certain behavior is infeasible and so give rise to perverse incentives or unintended undesirable consequences. For example, if the true reason such a private contingent contract is infeasible is that it violates an unspecified incentive constraint, then the uncontingent contract that is made contingent by the government policy also violates the same unspecified incentive constraint (for an early exposition of a version of this view, see Barro 1977).

These new models are often presented as essentially traditional Keynesian models derived from maximizing behavior, which has led to some confusion. Even though the labels IS and LM are often attached to certain equations, it is crucial to understand that these second-generation real business cycle models are built on the first-generation models, not on the Keynesian IS–LM model. That is, the New Keynesian models are simply real business cycle models with a few frictions added on. Thus, although it may be surprising to nonmacroeconomists, a canonical real business cycle model, augmented with money and flexible prices so that monetary policy can be meaningfully discussed, has essentially the same implications for the fraction of business cycle fluctuations explained by various shocks and, perhaps more surprisingly, the same implications for policy as a canonical New Keynesian model.

To see that classic real business cycle models and New Keynesian models both imply that technology shocks account for the vast bulk of fluctuations, consider two models. On the one hand, we have a classic real business cycle model by Prescott (1986). He compares the variance of detrended output in his model to the variance of detrended US output, and documents that 70 percent of the variance of the observed fluctuations in output in the US economy can be mechanically accounted for by productivity shocks (p. 16). On the other hand, using a state-of-the-art New Keynesian model, Justiniano, Primiceri, and Tambalotti (2010) find that technology shocks—here the sum of neutral and investment-specific technology shocks—account for 75 percent of the variance of output, which, somewhat surprisingly, is actually a higher percentage than that in Prescott's calculation.[1]

In sum, a typical New Keynesian model adds several frictions and shocks, but at its core, the key driving force for business cycles is a real business cycle model. Indeed, in the state-of-the-art New Keynesian model by Justiniano, Primiceri, and Tambalotti (2010), monetary policy shocks account for only a negligible fraction of the movements in output. In short, a New Keynesian model is exactly as Justiniano, Primiceri, and Tambalotti (p. 134) describe: "It is a medium-scale DSGE model with a neoclassical core, augmented with several frictions."

In part, the belief that real business cycle and New Keynesian models are based on different sources of economic fluctuations may represent a confusion about labeling. Some New Keynesian models like Smets and Wouters (2007) and Justiniano, Primiceri, and Tambalotti (2010) refer to investment-specific technology shocks as demand shocks, even though they represent shifts in the production function for the supply of investment goods, which might naturally seem to be types of supply shocks. Given the possibility for confusion on this point, these terms may have lost their usefulness.[2]

Moreover, under the popular narrative, New Keynesian models and flexible price models have radically different implications for monetary policy: in New Keynesian models, activist monetary policy is necessary to reduce the volatility of output and offset reductions in demand, whereas in flexible price models, monetary policy has no such role. We contend that this narrative reflects a deep misunderstanding of the workings and implications of these models. The genesis of this misunderstanding may be traced to the way New Keynesian models were presented, namely as traditional Keynesian models of the IS–LM variety but with maximizing agents.

This contention has been demonstrated by Correia, Nicolini, and Teles (2008), who show that the monetary and fiscal policy implications are identical in a flexible price model and a standard New Keynesian model with sticky prices. The flexible price model is a real business cycle model with essentially neutral money added on, in that money has little effect on output, in which consumers can purchase some goods with cash obtained in advance and some other goods with credit. Such a model is referred to as a cash-credit cash-in-advance model. The model features stochastic productivity shocks and stochastic government spending, as in Lucas and Stokey (1983), but is modified to incorporate differentiated varieties of a single

[1] We view this model as an updated version of that presented in the highly cited paper by Smets and Wouters (2007), which itself is a descendant of the model in the paper by Christiano, Eichenbaum, and Evans (2005). Briefly, Smets and Wouters (2007) exclude changes in inventories from their definition of investment and include the purchases of consumer durables in consumption rather than investment. Justiniano, Primiceri, and Tambalotti (2010), instead, include both the change in inventories and the purchases of consumer durables in investment. In other respects, the models are essentially identical.

[2] For another example where this terminology is less than helpful, consider a CES demand function for a differentiated product $y^d = (p/P)^{-\theta}Y$, where p is the price of that product, P is the aggregate price index, and Y is aggregate output. From the point of view of an individual producer, shifts in Y are shifts in that producer's demand curve, even when these shifts in Y come from aggregate productivity shocks. Here again, demand and supply terminology is not helpful.

consumption good sold by monopolistic competitors. The New Keynesian model is an identical model except that prices are sticky in that producers are allowed to adjust their prices only at random (Poisson) times. The set of instruments available to the government are the money supply (or equivalently, nominal interest rates) and state-contingent linear taxes on consumption and labor income.

The main result is that in both models, it is optimal to have identical policies: constant (producer) prices and tax rates set to smooth distortions by equating the relevant margins over time. Critically, if an outside observer had data from this sticky price economy under such an optimal policy, fluctuations in all aggregates would be identical to those generated by a frictionless real business cycle model, adjusted to include neutral money. The reason is that in the original sticky price economy, optimal monetary policy is not attempting to offset real shocks to the economy, but instead is attempting to reproduce the flexible price allocations of the frictionless version of the model (for related results, see Woodford 2003).

An immediate corollary of the work of Correia, Nicolini, and Teles (2008) is that the zero lower bound constraint, namely, the constraint that nominal interest rates cannot be negative, has no impact on the equivalence of policy in New Keynesian and flexible price models. Hence, when taxes are set optimally, the idea that the zero lower bound constraint makes increasing government spending especially attractive does not hold either (for details, see Correia, Farhi, Nicolini, and Teles 2013).

Finally, it is commonly argued that it is interesting to deprive the government of nearly all fiscal instruments when analyzing monetary policy because, in practice, it is difficult to quickly adjust fiscal policy in the depths of a recession. We argue that at least for deep recessions, this claim is not true: witness the speed at which the Obama stimulus program, formally, the American Recovery and Reinvestment Act of 2009, was passed. Regardless of the merits of this program, it was clearly passed quickly enough to affect the economic recovery.

In sum, New Keynesian models are most certainly not reincarnations of textbook IS–LM models with maximization added on. Rather, they are real business cycle models augmented with a few distortions—typically sticky prices and monopoly power—and shocks that do little to contribute to fluctuations or influence the nature of optimal policy.

Third Generation: Matching Aggregate Time Series Combined with External Validation

The goal of second-generation modern business cycle models, which were nearly all New Keynesian ones, was to help central banks in their medium-term forecasting and allow central banks to use them for counterfactual policy experiments in order to inform the policy debate. In contrast, the goal of third-generation models is to develop new and more deeply founded mechanisms that formalize alternative possible explanations for business cycles as well as provide convincing external validation for the quantitative importance of these newly formalized mechanisms.

Indeed, the hallmark of the third generation of modern business cycle models is their focus on an explicit external validation of their key mechanisms, using evidence independent of the particular aggregate time series for which the model is designed to account. Many of these third-generation models incorporate micro-level heterogeneity and are built on a tight connection between the mechanisms in the model and the wealth of micro-level data pertinent to the key forces in the model. A defining characteristic of these models, though, is neither the heterogeneity among agents in the model nor the micro-level evidence these models rely on, although both characteristics are common, but rather the insistence that any new parameters or feature that is included should be disciplined explicitly by direct evidence. In this sense, the spirit of the discipline of third-generation models echoes the attitudes of the original developers of first-generation models, except that third-generation models are sophisticated enough to match a wealth of additional aspects of the micro data and, in contrast to the first-generation models, do not need to be able to be aggregated to be solved.

More broadly, third-generation modern business cycle models grew naturally out of the first-generation ones. Only now, because of the development of sophisticated algorithms and the advent of high-powered computers, has it become feasible to explore third-generation models. Several decades ago, if a researcher was interested in a nonlinear model with both idiosyncratic and aggregate shocks, it was necessary to make assumptions so that the heterogeneity could be aggregated back to a suitably defined representative consumer and firm.

For example, in the classic model by Bernanke, Gertler, and Gilchrist (1999), even though banks are heterogeneous in their net worth, the model aggregates in that the only state variable of banks that needs to be recorded is aggregate net worth. The reason is that the model is carefully set up so that value functions are linear in net worth. With new algorithms and greater computing power, it is now feasible to compute such models even if they do not aggregate, so that the relevant aggregate state is the entire distribution of net worth across firms.

Many observers thought that the Great Recession would have led to an upheaval in macroeconomic modeling (for example, Christiano 2016). After all, these observers argued, much of the observed contraction in output was driven by disruptions in credit markets that spilled over into the real economy, but nearly all business cycle models featured no such links between financial and real activity. We argue that no upheaval in modeling has happened: in contrast to the Great Stagflation of the 1970s, the Great Recession has had essentially no impact on macroeconomic methodology per se. Rather, the Great Recession simply prompted macroeconomists to design models that elevated financial frictions from their previously modest role in amplifying the effects of other shocks, as in the classic work by Bernanke, Gertler, and Gilchrist (1999), to playing a central role in amplifying the shocks generating downturns.[3] The main

[3] A vibrant literature in international macroeconomics had already developed open economy models that included financial crises. However, the mechanisms explored in this work were not immediately applicable to the pattern of crises witnessed in large developed economies such as the United States.

consequence of the Great Recession was to push macroeconomists further away from the medium-scale New Keynesian models with hard-to-interpret shocks and frictions, chosen mainly for their ability to fit macro aggregates, and back to more elaborate versions of first-generation models of behavior modeled from primitives internally disciplined and externally validated by looking at their detailed implications for the data.

Although there are now many fine examples of third-generation modern business cycle models, below we discuss two examples of third-generation work with which we are most familiar.[4] In both examples, micro-level data are used to discipline the models' new features and to assess how the proposed mechanisms are borne out in the relevant data on consumers and firms. The illustration in the next section draws on the work of Arellano, Bai, and Kehoe (forthcoming), which is motivated both by micro-level and macro-level patterns of firm behavior and by the Great Recession of 2007–2009. The illustration in the following section focuses on our work in Kehoe, Midrigan, and Pastorino (forthcoming), which is motivated by the challenge for business cycle models to account for the cross-regional patterns of employment, consumption, and wages witnessed in the Great Recession.

Before reviewing these two examples, we compare the comovements of aggregates across the two largest postwar US recessions—the 1982 recession and the Great Recession—which helps explain the precise sense in which the Great Recession has been unusual.

Classifying and Modeling Recessions: 1982 and the Great Recession

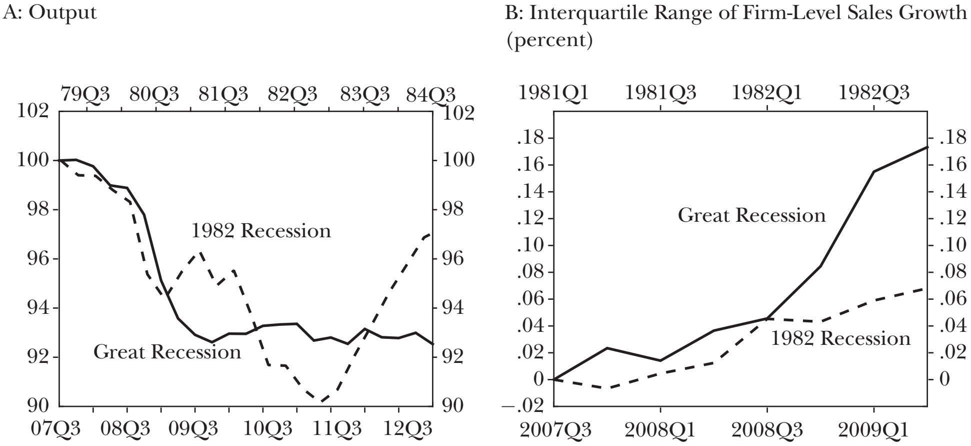
In terms of understanding and accounting for the Great Recession, two questions arise. First, can the Great Recession be thought of as a financial recession in a way that earlier large recessions such as the 1982 recession cannot be? Second, do the patterns of comovements between, say, output, hours, and productivity differ across financial and nonfinancial recessions?

To answer the first question, we draw on the work of Romer and Romer (2017), who argue that the 1982 recession in the United States exhibited no financial distress, whereas the Great Recession in the United States displayed some of the greatest financial distress in the entire post–World War II sample of developed

The patterns of crises in small emerging markets pointed to the central role of issues like a sudden stop of capital inflows (Mendoza 2010) and possible defaults on sovereign debt (like Cole and Kehoe 2000; Arellano 2008; Neumeyer and Perri 2005). These issues are clearly relevant to episodes in Argentina, Mexico, and Greece, but they played essentially no role in the US Great Recession.

[4] An important hybrid of second- and third-generation approaches is the work of Kaplan, Moll, and Violante (2018), which incorporates heterogeneous consumers into a New Keynesian model. On the one hand, whereas the costs of purchasing illiquid assets are disciplined by consumers' responses to unanticipated tax rebates, the key features of the model, namely the consumers' heterogeneous responses to monetary shocks, are not disciplined by the data. Moreover, computational limitations force the authors to consider only one-time unanticipated shocks, so that the implications of the model for business cycles are not yet known.

Figure 1
Comparing the 1982 Recession and the Great Recession



Note: In Figure 1A, output is indexed to 100 in 2007Q3.

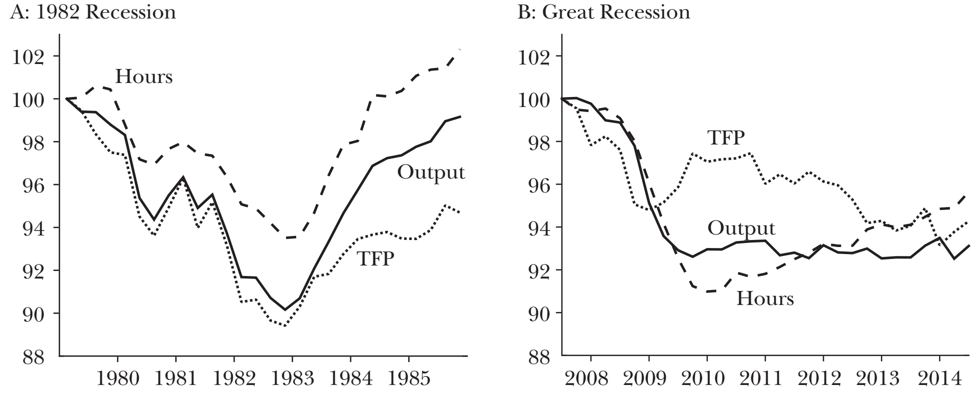
countries. Romer and Romer (2017) construct a financial distress measure based on a real-time narrative source, the *OECD Economic Outlook*, to identify the severity of a crisis by the size of the change of various indicators, including increases in the costs to financial institutions in obtaining funds and general increases in the perceived riskiness of financial institutions.

These authors show that throughout the entire 1982 recession, the distress measure for the United States indicated its lowest possible level—namely, no distress. Throughout the Great Recession, instead, this same measure indicates a growing level of distress that peaks in 2008 at a level indicating extreme financial distress. These two recessions are then clean cases to compare for the United States in terms of the comovements of macro aggregates, since they are the two deepest postwar recessions the country experienced, and they display the opposite extremes in the amount of financial distress.[5]

As panel A of Figure 1 shows, while the 1982 recession was somewhat deeper than the Great Recession, the downturn following the Great Recession was much

[5] In terms of modeling financial distress, an important issue is that of reverse causation. Regardless of the underlying cause, the deeper a recession is, the more likely that firms and households that contracted uncontingent loans will find it hard to repay them and, hence, the more likely that the financial institutions that extended such loans will experience financial distress. Moreover, the feedback is highly likely to go both ways: underlying causes, perhaps only partially financial, generate financial distress, and, in the presence of financial frictions, exacerbates the real downturn and leads to more distress. In short, an open question is whether the financial crisis is the result of a "shock" in the sense of an economy-wide run on financial institutions or whether financial frictions amplified other shocks and, hence, gave rise to severe financial distress.

Figure 2
US Output and Hours in the 1982 Recession and the Great Recession



Note: Panels A and B show output detrended by a 1.6 percent trend and non-detrended hours, normalized to 100 at the beginning of each period (1979Q1 for the 1982 recession and 2007Q3 for the Great Recession). TFP is total factor productivity.

more persistent.[6] The more basic question is whether the patterns of comovements among the major aggregates differ between financial and nonfinancial recessions to the point where a different mechanism is called for than those that conventionally account for most of the other postwar recessions. The short answer is "yes."

The comovements among output, hours, and total factor productivity in the Great Recession in the United States differed from earlier recessions. Compared to the 1982 recession, in the Great Recession the drop in total factor productivity was much smaller relative to the drop in output, whereas the drop in hours was much larger and longer-lasting than the drop in output. In terms of mechanisms, these patterns imply that the 1982 recession was characterized by the typical pattern of most postwar recessions, which can be mechanically accounted for by drops in total factor productivity, whereas the pattern in the Great Recession cannot be. This latter recession, instead, seems to suggest the need for a mechanism that makes labor fall much more relative to output than it does in both typical recessions and in standard models (Brinca, Chari, Kehoe, and McGrattan 2016).

As for the data, the two panels of Figure 2 illustrate this difference. In panel A, we graph output detrended by a 1.6 percent trend and non-detrended hours, both normalized to 100 in 1979Q1. We see that, relative to 1979Q1, output falls about 10 percent and hours fall about 6 percent so that the decline in hours is much smaller than the decline in output. In panel B, we graph output for the Great Recession detrended by a 1.6 percent trend, as well as non-detrended hours, both normalized to 100 in 2007Q3. Comparing the levels in 2007Q3 to those in the subsequent

[6] Some economists, such as Taylor (2016), argue that the causes of this slow growth are not directly connected to the financial crisis that accompanied the Great Recession.

trough, output falls about 7 percent and hours fall about 9 percent. Critically, during the Great Recession, the decline in hours is larger than the decline in output. Since standard real business models imply that for any given productivity shock, the percentage fall in hours is less than half of that in output, such models simply cannot account for the patterns of comovements in the Great Recession.

In sum, the 1982 recession, which exhibited no financial distress, was a typical real business cycle recession.[7] In contrast, the Great Recession, which exhibited financial distress an order of magnitude larger than all other postwar US recessions, had a modest fall in measured total factor productivity but a fall in hours greater than the fall in output.

A Mechanism for the Patterns of Comovements during the Great Recession

Any mechanism that accounts for the Great Recession must generate a large downturn in output associated with a sharp fall in hours, must generate a small decline in measured productivity, and must also be consistent with a large rise in measured financial distress.

One striking feature of the micro data from the Great Recession is that the financial crisis was accompanied by large increases in the cross-sectional dispersion of firm growth rates (Bloom, Floetotto, Jaimovich, Sparta-Eckstein, and Terry 2014). Indeed, as panel B of Figure 1 shows, the increase in the interquartile range of firms' sales growth during the Great Recession was nearly triple that during the 1982 recession. As credit conditions tightened during the financial crisis, firms' credit spreads increased while both equity payouts and debt purchases decreased. Motivated by these observations and the patterns of comovements described earlier, Arellano, Bai, and Kehoe (forthcoming) build a model with heterogeneous firms and financial frictions, in which increases in volatility at the firm level lead to increases in the cross-sectional dispersion of firm growth rates, a worsening of financial conditions, and a decrease in aggregate output and labor associated with small movements in measured total factor productivity.

The key idea in the model is that hiring inputs to produce output is risky: firms must hire inputs before they receive the revenues from their sales. To hire these inputs, firms must pledge to use some of their future revenues to pay for them. In this context, (owners of) firms face the risk of any idiosyncratic shock that occurs between the time of production and the receipt of revenues. When financial markets are incomplete in that firms have only access to debt contracts to insure against such shocks, firms and their creditors must bear this risk, which has real consequences if firms must experience a costly default once they cannot meet their financial obligations. In the model, an increase in uncertainty arising from an increase in the volatility of idiosyncratic productivity shocks at the firm level makes the revenues from any given amount of

[7] Chari, Kehoe, and McGrattan (2007) stress that measured fluctuations in total factor productivity are best thought of as efficiency wedges—namely, reduced-form shocks that arise from the interaction of frictions with primitive shocks. Hence, this finding could be consistent with the view that the decline in measured total factor productivity during the 1982 recession was a monetary policy reaction to nontechnology shocks.

labor hired more volatile and, thus, a default more likely. Thus, in equilibrium, an increase in volatility leads firms to hire fewer inputs, and so output to decrease.

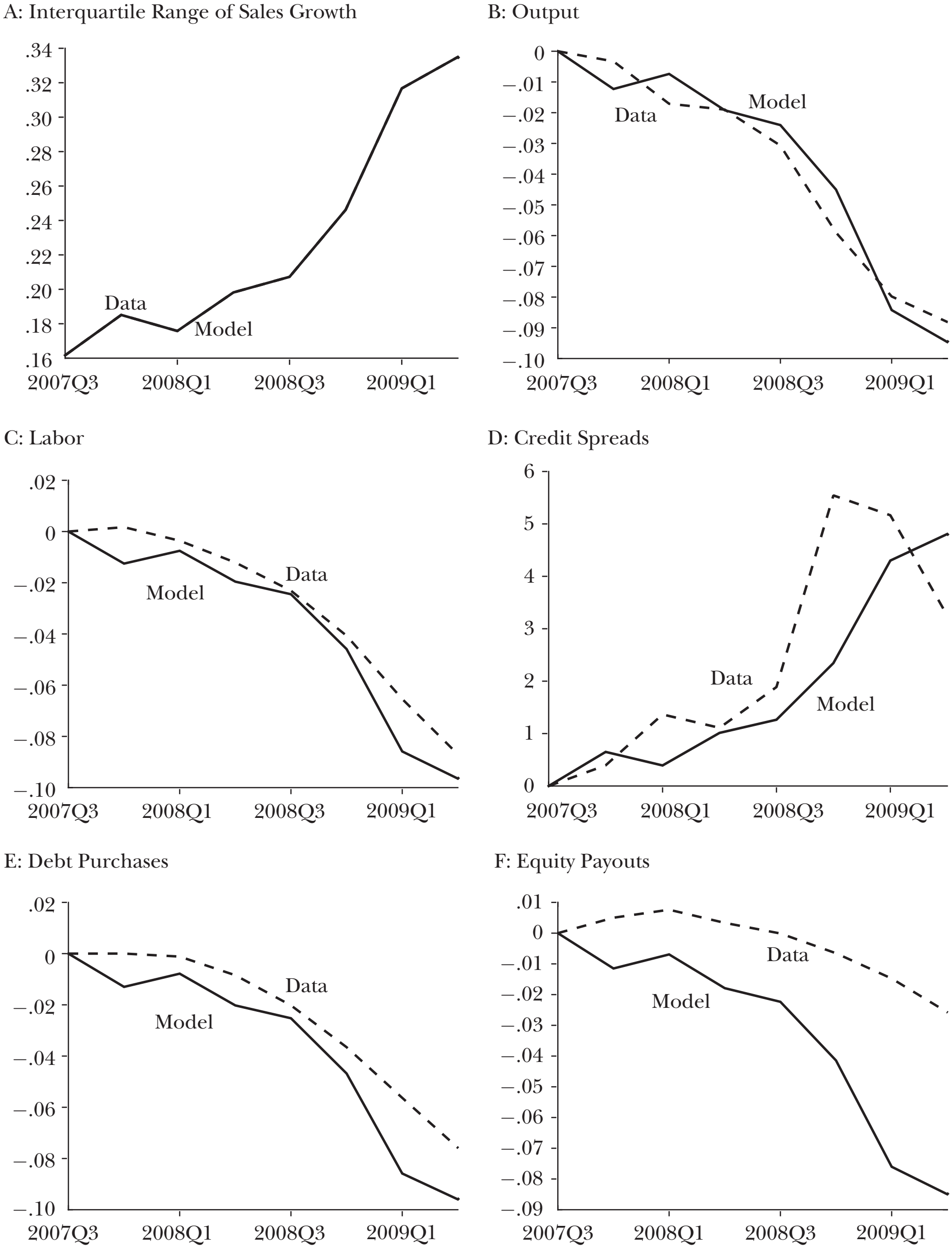
Formally, the model of Arellano, Bai, and Kehoe (forthcoming) features a continuum of heterogeneous firms that produce differentiated products. The productivity of these firms is subject to idiosyncratic shocks with stochastically time-varying volatility; these volatility shocks are the only aggregate shocks in the economy. Three ingredients are critical to the workings of the model. First, firms hire their inputs—here, labor—and produce before they know their idiosyncratic shocks. The insight that hiring labor is a risky investment is a hallmark of quantitative search and matching models, but is missing from most simple macroeconomic models. Second, financial markets are incomplete in that firms have access only to state-uncontingent debt and can default on it. Firms face interest rate schedules for borrowing that depend on all the shocks, so that higher borrowing and labor hiring result in higher probabilities of default. Third, motivated by the work of Jensen (1986), the model includes an agency friction in that managers can divert free cash flow to projects that benefit themselves at the expense of firms. This friction makes it optimal for firms to limit free cash flow and, thus, makes firms less able to self-insure against adverse shocks.

In the model, the main result is that an increase in uncertainty arising from an increase in the volatility of idiosyncratic productivity shocks increases the volatility of the revenues from any given amount of labor hired. As the risk of default increases, firms cut back on hiring inputs. This result depends critically on the assumptions of incomplete financial markets and the agency friction. If firms had access to complete financial markets, firms would simply respond to a rise in volatility by restructuring the pattern of payments across states and, as Arellano, Bai, and Kehoe (forthcoming) show, both output and labor would increase sharply when volatility rises. Indeed, when the distribution of idiosyncratic productivity spreads out and shocks are serially correlated, firms with high current productivity shocks tend to hire relatively more of the factor inputs. It is only when the volatility of firm-level productivity shocks is accompanied by financial frictions that the model produces a downturn. Without agency costs, firms could self-insure by maintaining a large buffer stock of unused credit. Absent the agency friction, firms find it optimal to build up buffer stocks well in excess of those observed in the data. With it, however, they find it optimal to limit the size of their buffer stocks and maintain debt levels consistent with those in the data. With such debt levels, the model generates realistic fluctuations in labor.

Quantitatively, Arellano, Bai, and Kehoe (forthcoming) investigate whether an increase in the volatility of firm-level idiosyncratic productivity shocks, which generates the increase in the cross-sectional dispersion of firm-level growth rates observed in the recent recession, leads to a sizable contraction in aggregate economic activity and tighter financial conditions. To do so, they choose a sequence of volatility shocks so that the model produces the same cross-sectional increase in sales growth as observed during the Great Recession. Figure 3A shows the resulting cross-sectional volatility of sales growth in the model and the data, where the latter is measured by the interquartile range of sales growth across firms. Figures 3B and 3C show that the model can account for essentially all of the contraction in output and labor

Figure 3

Great Recession Event: Data and a Model



Source: Arellano, Bai, and Kehoe (forthcoming).

Table 1

Financial Variables from the Cross-Sectional Distribution of Firms: Data and a Model

	Data (%)			*Model (%)*		
Percentile	25	50	75	25	50	75
Spread	1	1.3	2.1	1.1	2.8	6.3
Growth	−9	0	11	−7	0	9
Leverage	9	26	62	25	29	33
Debt purchases	−10	0	21	−14	0	16
Equity payouts	−4	0	12	−19	0	23

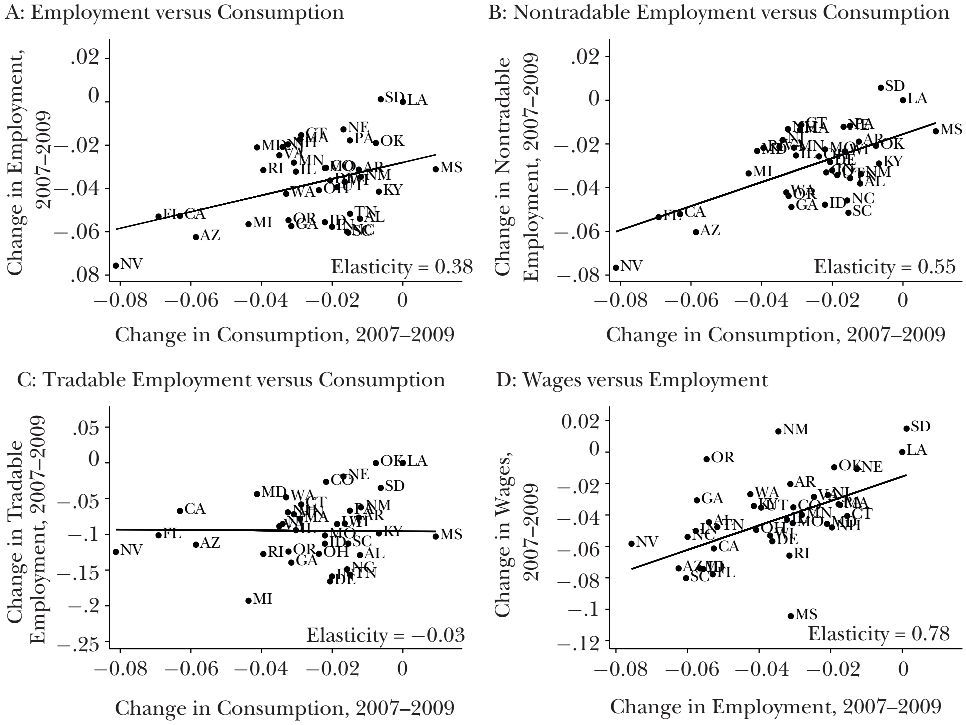
Source: The data are from Compustat, and the model is from Arellano, Bai, and Kehoe (forthcoming).
Note: Leverage is the sum of short-term and long-term debt divided by average sales. Equity payouts are the ratio of the sum of dividends and net equity repurchases to average sales and debt. Debt repurchases are the ratio of the change in total firm debt to average sales. For both data and model, we report the median of the time series of the 25th, 50th, and 75th percentiles across firms, computed for each variable and quarter. Growth and dividends are reported relative to the median 50th percentile.

that occurred in the Great Recession. Figures 3D, 3E, and 3F show that the model also does a reasonable job of reproducing the changes in financial variables that occurred during this period, as measured by credit spreads, debt purchases, and equity payouts. More generally, the authors show that their model generates labor fluctuations that are large relative to those in output, similar to the relationship between output and labor in the data.

It is useful to contrast the Arellano, Bai, and Kehoe (forthcoming) approach linking the model to the micro data to that used in a well-cited second-generation model of financial shocks and the Great Recession, namely, that of Christiano, Motto, and Rostagno (2017). This latter paper focuses on fitting 12 aggregate time series: GDP, consumption, investment, hours, inflation, wages, prices of investment goods, and the federal funds rate as well as four aggregate financial variables. The Christiano, Motto, and Rostagno (2017) paper represents the frontier work in that generation, but it never attempts to compare the detailed patterns of firm-level variables implied by the model to those in the data. In contrast, Arellano, Bai, and Kehoe (forthcoming) take a very different approach to the micro data. The authors start by showing that the model is consistent with some basic features of firms' financial conditions over the cycle, namely that firm credit spreads are countercyclical, as in the data, and that both the ratio of debt purchases to output and the ratio of equity payouts to output have similar correlations with output and volatility as in the data.

They then turn to micro moments of financial variables from the cross-sectional distribution of firms. Table 1 presents the time series median of spreads, the growth of sales, leverage, debt purchases, and equity payouts by each quartile. While there are some differences, this simple model does a reasonable job of reproducing these moments. Arellano, Bai, and Kehoe also show that the correlations of firm-level leverage with firm-level credit spreads, sales growth, debt purchases, and equity payouts are similar in the model and the data.

Figure 4
Change in Employment, Consumption, and Wages across US States


A: Employment versus Consumption

B: Nontradable Employment versus Consumption

C: Tradable Employment versus Consumption

D: Wages versus Employment

Source: Kehoe, Midrigan, and Pastorino (forthcoming).

A Mechanism for the Regional Patterns of the Great Recession

An alternative and complementary insight into the Great Recession can be gained by exploring the distinctively different characteristics of the Great Recession within different regions of the United States. To that end, we first discuss the regional patterns of employment, consumption, and wages in the United States during that time. We then conclude by presenting a promising mechanism that accounts for the strongly differential response of different US states to the Great Recession.

During the Great Recession, the regions of the United States that experienced the largest declines in household debt also experienced the largest drops in consumption, employment, and wages (for example, Mian and Sufi 2011, 2014). Here, we focus on two main aggregate patterns. First, the regions of the United States that were characterized by the largest declines in consumption were also characterized by the largest declines in employment, especially in the nontradable goods sector. Second, the regions that experienced the largest employment declines also experienced the largest declines in real wages relative to trend.

The panels of Figure 4 summarize these patterns. In Kehoe, Midrigan, and Pastorino (forthcoming), we illustrate the first pattern by using annual state-level data on employment and consumption from the Bureau of Economic Analysis. In the spirit of the model, we isolate changes in consumption associated with changes in households' ability to borrow—or, more generally, in credit conditions—as proxied by changes in house prices, by projecting state-level consumption growth on the corresponding growth in state-level house prices (from Zillow). We use the resulting series for consumption growth in our analysis (for a similar approach, see Charles, Hurst, and Notowidigdo 2015).

Panel A of Figure 4, taken from Kehoe, Midrigan, and Pastorino (forthcoming), plots state-level employment growth between 2007 and 2009 against the measure of state-level consumption growth just described over this same period. The elasticity of employment to consumption is 0.38. Panels B and C show that consumption declines are associated with relatively large declines in nontradable employment and essentially no changes in tradable employment across states: a 10 percent decline in consumption across states is associated with a 5.5 percent decline in nontradable employment and a negligible (and statistically insignificant) 0.3 percent increase in tradable employment. The large negative intercept in panel C shows that the decline in tradable employment is sizable in all states but unrelated to changes in consumption across states.

The second main correlation is shown in panel D of Figure 4, which reproduces a version of the findings in Beraja, Hurst, and Ospina (2016). These authors document that wages were moderately flexible in the cross section of US regions during the Great Recession: the cross-regional decline in wages was almost as large as the decline in employment. We closely follow their approach by using census data for wages from the Integrated Public Use Microdata Series and controlling for observable differences in workforce composition both across states and within a state over time, as in Beraja, Hurst, and Opsina (2016). As panel D shows, a decline in employment of 10 percent across US states during the Great Recession is associated with a decline in wages of 7.8 percent.

To investigate these cross-regional patterns, Beraja, Hurst, and Ospina (2016) use what they term a semi-structural methodology, which relies on a general equilibrium model and a combination of regional and aggregate data, to identify the regional and aggregate shocks driving business cycles. In particular, based on detailed census data at the household level on employment and wages, they find that, in the cross section, in regions where hours worked fell relatively more, nominal and real wages fall relatively more. These authors also show that shocks to the intertemporal marginal rate of substitution of consumption—called discount factor shocks—can account for the vast bulk of the cross-regional variation in employment in the United States during the Great Recession. The idea of using shocks to the discount factor as a proxy for variations in financial risk in the context of the Great Recession was also applied by Hall (2017).[8]

[8] Here we have discussed one class of models that accounts for aggregate movements and another one that accounts for cross-sectional movements. For an interesting model that attempts to account for both movements at the same time, see Jones, Midrigan, and Philippon (2017).

Using an approach that is complementary to Beraja, Hurst, and Ospina (2016), in Kehoe, Midrigan, and Pastorino (forthcoming), we investigate how the interplay between credit and labor market frictions can account for the cross-sectional patterns just documented. We develop a version of the Diamond–Mortensen–Pissarides search model with risk-averse agents, borrowing constraints, and human capital accumulation. The underlying idea is that hiring workers is an investment activity: costs of creating vacancies are paid up front, whereas benefits, as measured by the flows of surplus from the match between a firm and worker, accrue over time. In this framework, a credit tightening generates a fall in investment—including investment related to hiring workers—that induces firms to post fewer vacancies and so causes employment in the aggregate to fall.

The key innovation here is the addition of human capital accumulation on the job. In a textbook version of the Diamond–Mortensen–Pissarides search model without human capital accumulation, a large fraction of the present value of benefits from forming a match accrues early in the match. As a result, credit tightening has little effect on hiring in this model. But in the presence of human capital accumulation, the flows of benefits from forming a match have a much longer duration. Intuitively, a match not only produces current output but also augments a worker's human capital, which is also valuable to future matches and thus has persistent effects on a worker's output flows—a finding that holds even if matches dissolve at a high rate. We show that this significantly longer duration of surplus flows or returns to employment amplifies, by a factor of 10, the drop in employment from a credit contraction like the one observed during the Great Recession, relative to that implied by the model without human capital accumulation.

To build intuition for our new mechanism, consider a firm's incentives to post vacancies after a credit tightening that leads to a temporary fall in consumption. Since consumers have a desire to smooth consumption, this temporary fall in consumption increases consumers' marginal utility and hence their shadow price of current goods, which then mean-reverts. This temporary increase in the shadow price of goods has two opposing effects. First, it increases the cost of posting vacancies by raising the shadow value of the goods used in this investment. Second, it increases the surplus from a match by raising the shadow value of the surplus flows produced by a match. The cost of posting vacancies rises by more than the benefits because the cost of posting new vacancies is incurred immediately when goods are especially valuable, whereas, in the presence of human capital accumulation, the benefits accrue gradually in the future when shadow prices have already started to mean-revert. As a result, firms post fewer vacancies and, in the aggregate, employment contracts. The longer is the duration of the surplus flows from a match, the larger is the resulting drop in vacancies.

We show that the resulting model does an excellent job of reproducing the cross-state patterns of the Great Recession in terms of the comovement of consumption as well as nontradable, tradable, and overall employment. The model is also consistent with the observation that in the cross section of US states, wages are moderately flexible: a 10 percent drop in employment is associated with a fall in

Table 2
Individual Wages and Profits: Data and a Model

Moments	*Data*	*Model*
Targeted Moments		
Cross-sectional difference in log wages	1.21	1.19
30 to 1 years of experience		
Annual wage growth during an employment spell		
1–10 years of experience	0.10	0.10
11–20 years of experience	0.07	0.08
21–30 years of experience	0.06	0.06
31–40 years of experience	0.06	0.05
1–40 years of experience	0.07	0.07
Moments for External Validation		
Mean wage drop after nonemployment spell	0.044–0.055	0.05
Sensitivity of wage loss to one additional tenure year, %	1.54	1.95
Standard deviation of initial log wages	0.85	0.82
Profit share of revenue	0.06	0.06

Note: For details, see Kehoe, Midrigan, and Pastorino (forthcoming).

wages of 7.8 percent in both the data and the model. Thus, the model predicts sizable employment changes in response to a credit tightening, even though wages are as flexible as they are in the data. As Beraja, Hurst, and Ospina (2016) emphasize, this finding of substantial wage flexibility in the data casts doubt on the popular explanations of the Great Recession in the New Keynesian literature.

It is helpful to contrast second- and third-generation modern business cycle model approaches to understand the cross-regional features of the Great Recession discussed above. The second-generation approach would simply imply choosing parameters for the human capital process so as to fit the state-level employment patterns observed in the data, without informing this choice with any specific micro evidence on the relationship between human capital accumulation and wage growth or verifying whether the inferred parameters are consistent with additional micro evidence.

Instead, we proceed as follows. Because the process for human capital accumulation is critical for the model's predictions, we take great care in using micro data to quantify it. The top part of Table 2 illustrates how we use cross-sectional wage differences from Elsby and Shapiro (2012) to learn how wages vary with experience, as well as longitudinal data on how wages grow over an employment spell, from Buchinsky, Fougère, Kramarz, and Tchernis (2010), to discipline the model parameters.

The bottom part of Table 2 shows how we used other evidence from the micro data, not directly targeted in our moment-matching exercise, for external validation of our mechanism. We show that the model reproduces well observed drops in wages after a nonemployment spell, the sensitivity of this wage drop to an additional year of tenure on the job, the standard deviation of wages at the beginning of an employment spell, and the profit share of revenue. The model is also consistent with other patterns, including the distribution of durations of nonemployment spells and the evidence on wage losses from displaced worker regressions (as in Jacobson, LaLonde, and Sullivan 1993).

Finally, we show that our main result on the employment decline in response to a credit tightening is robust to a range of estimates of wage growth in the labor economics literature.

Thus, this third-generation real business cycle model introduces a new mechanism, human capital accumulation, for the amplification of the employment response to a credit crunch, and does so in a way that is disciplined by evidence that is external to the phenomenon to be explained.

Conclusion: The Centrality of Shifts in Method

The real business cycle revolution, at its core, was a revolution of method. It represents a move from an older econometric methodology underlying traditional Keynesian and monetarist large-scale macroeconomic models, in which exclusionary restrictions in a system of equations were taken to be the primitive specification of behavior, toward an approach in which maximization problems for consumers and firms that are consistent with a notion of general equilibrium are taken to be the primitive specification of behavior.

It is most fruitful to think of this methodology as a highly flexible language through which modern macroeconomists communicate. The class of existing real business cycle models using dynamic stochastic general equilibrium methods has come to include an enormous variety of work: real and monetary; flexible price and sticky price; financial and labor market frictions; closed and open economies; infinitely lived consumer and overlapping generations versions; homogeneous agent and heterogeneous agent versions; rational and robust expectations; time inconsistency issues at either the policymaker level or the individual decision maker level; multiple equilibria, constrained efficient equilibria, and constrained inefficient equilibria; coordination failures; and so on. Indeed, the language seems flexible enough to incorporate any well-thought-out idea.

What distinguishes individual papers that adopt this language, then, is not the broad methodology used, but rather the particular questions addressed and the specific mechanisms built into the model economy. For example, if one is interested in investigating optimal monetary policy in the face of financial shocks to the credit system, it is necessary to model monetary policy, financial shocks, and a credit system. But in every case, the unifying feature of real business cycles is their methodology—the specification of primitive technology, preferences, information structure, and constraints in an environment in which agents act in their own interest.

Macroeconomists still have fundamental disputes, but they all revolve around methodology. In particular, some maintain that all restrictions on prices, wages, and contracts must arise from economic fundamentals, such as technologies, including commitment technologies, preferences, and information structure. For these macroeconomists, the existing sticky wage and sticky price models are unappealing because, as Barro (1977) explained, even if wages and prices are sticky in that they cannot respond to shocks, there typically are feasible and mutually beneficial contracts that dominate them. Once such contracts are adopted, the case for an activist monetary policy is strongly weakened. Such macroeconomists also find unappealing models in which debt contracts cannot depend on aggregate observable variables, such as output or region-wide house prices, even though these variables are outside the ability of any single agent to affect, so no moral hazard issue would arise if contracts depended on them. In these setups, they find particularly methodology the study of policies that simply allow the government to partially replicate outcomes that private agents should be able to achieve naturally by themselves.

More important, although macroeconomists often hold heterogeneous beliefs about how promising any particular mechanism may be in accounting for features of the data or about the benefits of any particular policy, they agree that a disciplined debate rests on communication in the language of dynamic general equilibrium theory. By so doing, macroeconomists can clarify the origins of any disagreement and hence make progress on how to settle it. For example, when two different views are justified by fully specified quantitative models, it is relatively easy to pinpoint which key parameters or mechanisms are at the heart of the differing conclusions for policy. Hence, future work can attempt to discern which is in greater conformity with the data. In sum, this approach turns disagreements about outcomes of policies, which are difficult to make scientific progress on without a model, into disagreements about fundamental parameters, which are easier to resolve.

In this sense, there is no crisis in macroeconomics, no massive failure in methodology, no need for undisciplined frictions and shocks. Overall, modern macroeconomists live under a big tent that welcomes creative ideas laid out in a coherent language, specified at the level of primitives, and disciplined by external validation.

■ *We thank Adrien Auclert, Mark Gertler, Robert Hall, Gordon Hanson, Robert Lucas, Ellen McGrattan, Juan Pablo Nicolini, John Taylor, and Timothy Taylor for their comments and are especially thankful to John Cochrane for his detailed feedback. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.*

References

Arellano, Cristina. 2008. "Default Risk and Income Fluctuations in Emerging Economies." *American Economic Review* 98(3): 690–712.

Arellano, Cristina, Yan Bai, and Patrick J. Kehoe. Forthcoming. "Financial Frictions and Fluctuations in Volatility." *Journal of Political Economy.*

Barro, Robert J. 1977. "Long-Term Contracting, Sticky Prices, and Monetary Policy." *Journal of Monetary Economics* 3(3): 305–16.

Beraja, Martin, Erik Hurst, and Juan Ospina. 2016. "The Aggregate Implications of Regional Business Cycles." NBER Working Paper 21956.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." Chap. 21 in *Handbook of Macroeconomics*, vol. 1C, edited by John B. Taylor and Michael Woodford, 1341–93. Amsterdam: North Holland, Elsevier.

Bloom, Nicholas, Max Floetotto, Nir Jaimovich, Itay Saporta-Eksten, and Stephen Terry. 2014. "Really Uncertain Business Cycles." Unpublished paper.

Brinca, Pedro, V. V. Chari, Patrick J. Kehoe, and Ellen R. McGrattan. 2016. "Accounting for Business Cycles." Chap. 14 in *Handbook of Macroeconomics*, vol. 2A, edited by John B. Taylor and Harald Uhlig, 1013–63. Amsterdam: North Holland, Elsevier.

Buchinsky, Moshe, Denis Fougère, Francis Kramarz, and Rusty Tchernis. 2010. "Interfirm Mobility, Wages and the Returns to Seniority and Experience in the United States." *Review of Economic Studies* 77(3): 972–1001.

Chari, V. V., Patrick J. Kehoe, and Ellen R. McGrattan. 2007. "Business Cycle Accounting." *Econometrica* 75(3): 781–836.

Charles, Kerwin K., Erik Hurst, and Matthew J. Notowidigdo. 2015. "Housing Booms and Busts, Labor Market Opportunities, and College Attendance." NBER Working Paper 21587.

Christiano, Lawrence J. 2016. "The Great Recession: Earthquake for Macroeconomics." *Macroeconomic Review,* Monetary Authority of Singapore 15(1): 87–96.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *Journal of Political Economy* 113(1): 1–45.

Christiano, Lawrence J., Roberto Motto, and Massimo Rostagno. 2017. "Risk Shocks." *American Economic Review* 104(1): 27–65.

Cole, Harold L., and Timothy J. Kehoe. 2000. "Self-Fulfilling Debt Crises." *Review of Economic Studies* 67(1): 91–116.

Cooper, Russell, and João Ejarque. 2000. "Financial Intermediation and Aggregate Fluctuations: A Quantitative Analysis." *Macroeconomic Dynamics* 4(4): 423–47.

Correia, Isabel, Emmanuel Farhi, Juan Pablo Nicolini, and Pedro Teles. 2013. "Unconventional Fiscal Policy at the Zero Bound." *American Economic Review* 103(4): 1172–1211.

Correia, Isabel, Juan Pablo Nicolini, and Pedro Teles. 2008. "Optimal Fiscal and Monetary Policy: Equivalence Results." *Journal of Political Economy* 116(1): 141–70.

Elsby, Michael W. L., and Matthew D. Shapiro. 2012. "Why Does Trend Growth Affect Equilibrium Employment? A New Explanation of an Old Puzzle." *American Economic Review* 102(4): 1378–1413.

Hall, Robert E. 2017. "High Discounts and High Unemployment." *American Economic Review* 107(2): 305–30.

Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan. 1993. "Earnings Losses of Displaced Workers." *American Economic Review* 83(4): 685–709.

Jensen, Michael C. 1986. "The Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers." *American Economic Review* 76(2): 323–30.

Jones, Callum, Virgilliu Midrigan, and Thomas Philippon. 2017. "Household Leverage and the Recession." Unpublished paper.

Justiniano, Alejandro, Giorgio E. Primiceri, and Andrea Tambalotti. 2010. "Investment Shocks and Business Cycles." *Journal of Monetary Economics* 57(2): 132–45.

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante. 2018. "Monetary Policy According to HANK." *American Economic Review* 108(3): 697–743.

Kehoe, Patrick J., Virgiliu Midrigan, and Elena Pastorino. Forthcoming. "Debt Constraints and Employment." *Journal of Political Economy***.**

Kydland, Finn E., and Edward C. Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50(6): 1345–70.

Lagos, Ricardo. 2006. "A Model of TFP." *Review of Economic Studies* 73(4): 983–1007.

Long, John B., Jr., and Charles I. Plosser. 1983. "Real Business Cycles." *Journal of Political Economy* 91(1): 39–69.

Lucas, Robert E., Jr. 1976. "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy*, vol. 1, edited by Karl Brunner and Allan H. Meltzer, 19–46. Amsterdam: North-Holland.

Lucas, Robert E., Jr., and Nancy L. Stokey. 1983. "Optimal Fiscal and Monetary Policy in an Economy without Capital." *Journal of Monetary Economics* 12(1): 55–93.

Mendoza, Enrique G. 2010. "Sudden Stops, Financial Crises, and Leverage." *American Economic Review* 100(5): 1941–66.

Mian, Atif, and Amir Sufi. 2011. "House Prices, Home Equity–Based Borrowing, and the US Household Leverage Crisis." *American Economic Review* 101(5): 2132–56.

Mian, Atif, and Amir Sufi. 2014. "What Explains the 2007–2009 Drop in Employment?" *Econometrica* 82(6): 2197–2223.

Neumeyer, Pablo A., and Fabrizio Perri. 2005. "Business Cycles in Emerging Economies: The Role of Interest Rates." *Journal of Monetary Economics* 52(2): 345–80.

Prescott, Edward C. 1986. "Theory Ahead of Business Cycle Measurement." *Federal Reserve Bank of Minneapolis Quarterly Review* 10(4): 9–22.

Romer, Christina D., and David H. Romer. 2017. "New Evidence on the Aftermath of Financial Crises in Advanced Countries." *American Economic Review* 107(10): 3072–118.

Romer, Paul. 2016. "The Trouble with Macroeconomics." Delivered January 5, 2016, as the Commons Memorial Lecture of the Omicron Delta Epsilon Society. Forthcoming in *The American Economist*.

Sargent, Thomas J., and Neil Wallace. 1975. "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *Journal of Political Economy* 83(2): 241–54.

Smets, Frank, and Rafael Wouters. 2007. "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach." *American Economic Review* 97(3): 586–606.

Summers, Lawrence H. 1986. "Some Skeptical Observations on Real Business Cycle Theory." *Federal Reserve Bank of Minneapolis Quarterly Review* 10(4): 23–27.

Taylor, John B. 2016. "Slow Economic Growth as a Phase in a Policy Performance Cycle." *Journal of Policy Modeling* 38(4): 649–55.

Woodford, Michael. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 167–194

Microeconomic Heterogeneity and Macroeconomic Shocks

Greg Kaplan and Giovanni L. Violante

In this essay, we discuss the emerging literature in macroeconomics that combines heterogeneous agent models, nominal rigidities, and aggregate shocks. This literature opens the door to the analysis of distributional issues, economic fluctuations, and stabilization policies—all within the same framework.

Quantitative macroeconomic models have integrated heterogeneous agents and incomplete markets for nearly three decades, but they have been mainly used for the investigation of consumption and saving behavior, inequality, redistributive policies, economic mobility, and other cross-sectional phenomena. Representative agent models have remained the benchmark in the study of aggregate fluctuations (for reasons we will discuss later). However, the Great Recession bluntly exposed the shortcomings of a representative-agent approach to business cycle analysis. A broadly shared interpretation of the Great Recession places its origins in housing and credit markets. The collapse in house prices affected households differently, depending on the composition of their balance sheets. The extent to which this

■ *Greg Kaplan is Professor of Economics, University of Chicago, Chicago, Illinois; Research Fellow, Institute for Fiscal Studies, London, United Kingdom; and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Giovanni L. Violante is Professor of Economics at Princeton University, Princeton, New Jersey; Research Fellow, Centre for Economic Policy Research, London, United Kingdom; International Research Fellow, Institute for Fiscal Studies, London, United Kingdom; Research Fellow, Institute of Labor Economics (IZA), Bonn, Germany; and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are gkaplan@uchicago.edu and violante@princeton.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.167 doi=10.1257/jep.32.3.167

wealth destruction translated into a fall in expenditures was determined by marginal propensities to consume, which are also very heterogeneous and closely related to households' access to liquidity (Mian, Rao, and Su 2013; Kaplan, Violante, and Weidner 2014). Finally, this drop in aggregate consumer demand and the contemporaneous breakdown in bank lending to businesses (as explained by Gertler and Gilchrist in this issue) resulted in a severe contraction of labor demand which materialized unevenly across different occupations and skill levels. All this took place against the backdrop of a secular rise in income and wealth inequality.

Thus, portfolio composition, credit, liquidity, marginal propensities to consume, unemployment risk, and inequality were all central to the unfolding of the Great Recession. Yet these are all issues that one cannot discuss in a representative agent model (at least not without trivializing them). Indeed, the need for macroeconomists to move beyond the representative agent fiction in business cycle analysis was also emphasized by a number of high officials and governors of central banks in speeches delivered after the crisis, including Janet Yellen (2016) of the US Federal Reserve, Vitor Costancio (2017) of the European Central Bank, and Haruiko Kuroda (2017) of the Bank of Japan.

In response to these limitations of the representative agent approach to economic fluctuations, a new framework has emerged that combines key features of heterogeneous agents (HA) and New Keynesian (NK) economies. These HANK models offer a much more accurate representation of household consumption behavior and can generate realistic distributions of income, wealth, and, albeit to a lesser degree, household balance sheets. At the same time, they can accommodate many sources of macroeconomic fluctuations, including those driven by aggregate demand. In sum, they provide a rich theoretical framework for quantitative analysis of the interaction between cross-sectional distributions and aggregate dynamics.

In this article, we outline a state-of-the-art version of HANK based on Kaplan, Moll, and Violante (2018), together with its representative agent counterpart. We use this HANK model, calibrated to the US economy, to convey two broad messages about the role of household heterogeneity for the response of the macroeconomy to aggregate shocks.

The first message is that the similarity between the Representative Agent New Keynesian (RANK) and HANK frameworks depends crucially on the shock being analyzed. We illustrate this point through a series of examples. In response to a demand shock arising from a change in household discount factors, HANK and RANK generate the same aggregate dynamics through largely the same economic mechanisms. In response to a technology shock, the two models also generate similar aggregate dynamics but through different economic mechanisms. And following fiscal and monetary policy shocks, the two models generate different aggregate responses. These discrepancies can be traced to the fact that household consumption is more sensitive to income and less sensitive to interest rates in heterogeneous agent models than in representative agent models.

We then turn to our second message: certain important macroeconomic questions concerning economic fluctuations can only be addressed within heterogeneous agent models. To make this point, we look at how, in HANK models, aggregate demand shocks can have a proper microfoundation: for example, through unexpected changes in borrowing capacity or in uninsurable income risk. We also show how one can learn about the source and transmission mechanism of aggregate shocks by examining how they impact households at different parts of the wealth distribution. Finally, we illustrate how HANK models can be used to understand the effect of aggregate shocks and stabilization policies on household inequality.

We conclude by suggesting several broad directions for the future development of HANK models. Throughout this article, we focus on *household* heterogeneity, so when we use the term "agents" we refer to "households." There is a parallel literature on firm heterogeneity and aggregate dynamics, which deserves its own separate treatment.[1] Here, it suffices to say that many of the points we make on the role of heterogeneity apply to that literature as well.

Heterogeneity and Business Cycles in Macroeconomics, So Far

Macroeconomics is about general equilibrium analysis. Dealing with distributions while at the same time respecting the aggregate consistency dictated by equilibrium conditions can be extremely challenging. This explains why in the 1970s, when the path-breaking work of James Heckman and Daniel McFadden was paving the way for a rich treatment of cross-sectional heterogeneity in microeconometrics, the focus in macroeconomics was on developing models where aggregate outcomes would not depend on distributions. At that time, James Tobin famously defined macroeconomics as a subject that attains workable approximations by ignoring effects on aggregates of distributions of wealth and income (Sargent 2015). Heterogeneity was neutralized by assuming either identical initial conditions right off-the-bat, or special preference specifications (through Gorman aggregation), or complete markets (through Negishi aggregation).

Heterogeneous agent incomplete-markets models with nontrivial distributions of households appeared in the mid 1980s. Ljungqvist and Sargent (2004) baptized this class of models "Bewley models" because Truman Bewley (1983) was the first to explore the equilibrium properties of these economies. Throughout the 1990s, the seminal work of Imrohoroğlu (1989), Huggett (1993), Aiyagari (1994), and Ríos-Rull (1995), among others, laid the foundations for a new workhorse of quantitative macroeconomics that expanded the Bewley model and recast it in the recursive language developed by Robert Lucas, Edward Prescott, Thomas Sargent, and Nancy Stokey, among others. To quote from Aiyagari (1994, p. 1), its distinctive feature was that "aggregate behavior is the result of market interaction among a large number

[1] For example, see Caballero (1999) and Khan and Thomas (2008) for the debate on how firm-level nonconvex adjustment costs influence aggregate investment and Gertler and Gilchrist (1994) and Ottonello and Winberry (2018) for the debate on how firm-level financial constraints affect the transmission of monetary policy.

of agents subject to idiosyncratic shocks. ... This contrasts with representative agent models where individual dynamics ... coincide with aggregate dynamics ...”

This framework combines two building blocks. On the production side, a representative firm with a neoclassical production function rents capital and labor from households to produce a final good. On the household side, a continuum of agents each solve their own income fluctuation problem—the problem of how to smooth consumption when income is subject to random shocks and the only available financial instrument is saving (and possibly limited borrowing) in a risk-free asset (for example, Schechtman 1976). The equilibrium real interest rate is determined by equating households’ supply of savings to firms’ demand for capital.

The main motivation for modeling consumer behavior along these lines was the rapidly mounting empirical evidence, based on longitudinal household survey data, that most households fail in their efforts to perfectly smooth consumption (Hall 1978; Cochrane 1991; Attanasio and Davis 1996), a finding that time has only reinforced. Heterogeneous agent models allowed investigation of imperfect consumption insurance—its extent, reasons, and effects for the macroeconomy.

Reading through the recent surveys of this literature (for example, Heathcote, Storesletten, and Violante 2009; Guvenen 2011; Quadrini and Ríos-Rull 2015; Benhabib and Bisin forthcoming; De Nardi and Fella 2017), one is struck by the fact that while heterogeneous agent models have been routinely used to study questions pertaining to income and wealth inequality, redistribution, economic mobility, and tax reforms, until recently they had not been much used to study business cycles. The reason, we think, is twofold: computational complexity and a result known as “approximate aggregation.”

Computational complexity arises because in the recursive formulation of heterogeneous agent models with aggregate shocks, households require a lot of information in order to solve their dynamic optimization problems: each household must not only know its own place in the cross-sectional distribution of income and wealth, but must also understand the equilibrium law of motion for the entire wealth distribution. Under rational expectations, this law of motion is an endogenous equilibrium object, and solving for it is a computationally intensive process.

The first to successfully tackle this challenge were Krusell and Smith (1998), who pioneered the most well-known method and applied it to a simple heterogeneous agent economy with aggregate technology shocks. Despite recent advances in computing power and numerical methods, applying their method to the most interesting versions of heterogeneous agent economies remains challenging. In recent years, several new computational methods have been proposed that have widened the set of models that can be accurately solved. These include mixtures of projection and perturbation (Reiter 2009), mixtures of finite difference methods and perturbation (Ahn, Kaplan, Moll, Winberry, and Wolf 2017), adaptive sparse grids (Brumm and Scheidegger 2017), polynomial chaos expansions (Pröhl 2017), machine learning (Duarte 2018; Fernández-Villaverde, Hurtado, and Nuño 2018), and linearization with impulse-response functions (Boppart, Krussel, and Mitman 2017). Which of these, or other, methods will ultimately prevail is an open question.

The “approximate aggregation” result, uncovered by Krusell and Smith (1998), states that in many heterogeneous agent models, the mean of the equilibrium wealth distribution is sufficient to forecast all relevant future prices. The underlying logic is compelling: what matters for the aggregate dynamics of interest rates are the actions of households who hold the bulk of the wealth in the economy. Those rich households are well-insured against fluctuations and have saving functions that are approximately linear in wealth. Households that are close to the borrowing constraint, where the saving functions have curvature, are largely irrelevant in terms of their contribution to the aggregate capital stock and consumption. Krusell and Smith showed that in a benchmark version of the heterogeneous agent model, the aggregate dynamics of output, consumption, and investment in response to a shock to total factor productivity are almost identical to their counterpart representative agent model.

Approximate aggregation has proved surprisingly robust over time and has led many economists to conclude that aggregate dynamics in representative and heterogeneous agent models are essentially equivalent. As we show in this article, this interpretation of the original Krusell–Smith insight is inaccurate. Because of this misunderstanding, deviating from the representative agent approach was perceived by much of the profession as incurring a high computational cost for only little benefit. As a consequence, quantitative heterogeneous agent models rarely crossed paths with the study of business cycles.

The Great Recession put consumption, income, and wealth distributions back at the center stage of business cycles analysis and undermined this perception. Economists began to realize that two critical ingredients were needed for a coherent analysis of fluctuations and stabilization policy: 1) household heterogeneity; and 2) a framework that can accommodate aggregate demand shortfalls. In response, a number of macro researchers chose to address this gap in the most natural way: by combining key features of heterogeneous agent models and New Keynesian models.

Heterogeneous Agent New Keynesian (HANK) Models

In this section, we first argue that modeling household heterogeneity is important, by itself and in conjunction with nominal rigidities. Next, we discuss some early applications of HANK models. Finally, we outline this new framework in detail, setting the stage for the second part of our article where we contrast HANK and RANK models.

Heterogeneity is Key for Matching Facts about Consumption Behavior

Consumption behavior in representative agent models is inconsistent with empirical evidence. A representative household is essentially a permanent-income consumer facing an intertemporal budget constraint. As such, its consumption is extremely responsive to changes in current and future interest rates but barely responds to transitory changes in income.

The high sensitivity of consumption to interest rates is not well supported by macro or micro data. Analyses using aggregate time-series data typically find that, after controlling for aggregate income, consumption is not very responsive to changes in interest rates (Deaton 1987; Campbell and Mankiw 1989; Yogo 2004; Canzoneri, Cumby, and Diba 2007). A number of studies reveal that both the sign and size of the effect of changes in interest rates on consumption depend on households' net asset positions (Flodén, Kilström, Sigurdsson, and Vestman 2016; Cloyne, Ferreira, and Surico 2016). Empirical analyses using micro data on household portfolios also conclude that a sizable fraction of households (around one-third in the United States) hold close to zero liquid wealth or are near their borrowing limits (Kaplan, Violante, and Weidner 2014). Empirically, these households do not react to movements in interest rates (Vissing-Jørgensen 2002).

The implication from a representative agent model that consumption is insensitive to transitory income shocks is also inconsistent with the vast micro empirical literature surveyed by Jappelli and Pistaferri (2010). This literature has employed three approaches to identify exogenous income shocks. The first approach seeks quasi-experimental settings where natural variation generates randomness in either the receipt, amount, or timing of gains or losses. Examples include firm-level shocks, unemployment due to plant closings, stimulus payments and lottery winnings (for example, Browning and Crossley 2001; Johnson, Parker, and Souleles 2006; Broda and Parker 2014; Misra and Surico 2014; Fagereng, Holm, and Natvik 2016; Baker forthcoming). The second approach extracts the consumption response to the transitory component of regular income fluctuations by assuming a particular statistical process for income and exploiting assumptions about how income and consumption should co-vary (for example, Blundell, Pistaferri, and Preston 2008; Heathcote, Storesletten, and Violante 2014; Kaplan, Violante, and Weidner 2014). The third approach uses survey questions that ask respondents about how their expenditures would change in response to actual or hypothetical changes in their budgets (for example, Shapiro and Slemrod 2003; Christelis, Georgarakos, Jappelli, Pistaferri, and van Rooij 2017; Fuster, Kaplan, and Zafar 2018).

This collective body of evidence on marginal propensities to consume (MPCs) points towards: 1) sizable average MPCs out of small, unanticipated, transitory income changes; 2) larger MPCs for negative than for positive income shocks; 3) small MPCs in response to announcements about future income gains; and 4) substantial heterogeneity in MPCs that is correlated with access to liquidity. None of these four features are in line with the consumption behavior in representative agent models.

Heterogeneous agent models with incomplete markets can, instead, reproduce many of these aspects of consumption behavior. Households who are at a kink in their budget sets (generated, for example, by a borrowing limit or by a wedge between interest rates on liquid savings and unsecured borrowing) have high MPC out of transitory income shocks and do not respond to small changes in interest rates. For households who are close to a kink, exposure to uninsurable income risk raises the possibility of hitting the kink in the future, which shortens their effective time horizon, dampens the intertemporal substitution channel and raises their MPC (Carroll 1997). For all other households, a fall in real rates leads to an increase in expenditures through intertemporal substitution, but there is also a counteracting income effect that can be especially strong for wealthy households.

Heterogeneity Restores Keynesian Insights into the New Keynesian Model

During the last couple of decades, the New Keynesian model has become the reference paradigm for economists working for central banks and governments who needed a micro-founded framework to think about the aggregate and welfare effects of fiscal and monetary policy interventions (Clarida, Gali, and Gertler 1999; Woodford 2003). In a New Keynesian model, monopolistically competitive firms produce differentiated goods and face costs of adjusting prices. Because prices are sticky in the short-run, money supply can affect aggregate demand and monetary policy can have real effects. Over time, this research program has given rise to large-scale models that can accommodate multiple real and nominal aggregate shocks.

However, since the baseline New Keynesian model employs a representative agent, its implied consumption dynamics feature strong intertemporal substitution and weak income sensitivity. Thus, somewhat paradoxically and in spite of its name, the mechanism by which aggregate demand affects aggregate output in the standard New Keynesian model differs markedly from the ideas typically associated with John Maynard Keynes (namely, the equilibrium spending multiplier). For these reasons, Cochrane (2015) has suggested that it would be more appropriate to call this model the "sticky-price intertemporal substitution model."

Relative to the representative agent version, the heterogeneous agent version of the New Keynesian model has a higher average MPC, a more realistic distribution of MPCs, and a lower sensitivity to interest rates, which makes the general equilibrium effects of aggregate demand fluctuations much more salient in the heterogeneous agent version.

HANK: Early Examples

The first examples of heterogeneous agent New Keynesian models appeared in the immediate wake of the Great Recession. These models were designed to address the origins of the crisis, its propagation, and the observed policy responses, all aspects in which household heterogeneity in terms of income, wealth, and balance sheets plays a central role. Oh and Reis (2012) study the extent to which fiscal stimulus in the form of targeted transfers to households alleviated the costs of the recession. Guerrieri and Lorenzoni (2017) examine the impact of a tightening of household borrowing constraints on aggregate demand and output. McKay and Reis (2016) investigate the role of automatic stabilizers in dampening macroeconomic fluctuations when monetary policy is active and when it is constrained by the zero lower bound. Similarly, Krueger, Mitman, and Peri (2016) examine the effectiveness of unemployment insurance in mitigating the fall in aggregate expenditures during the crisis. McKay, Nakamura, and Steinsson (2016) and Werning (2015) study the effectiveness of various forms of monetary policy including forward guidance, an instrument used by central banks to stimulate aggregate demand when the zero lower bound is binding. We also study this in Kaplan, Moll, and Violante (2016). Den Haan, Rendahl, and Riegler (2017) and Bayer, Lütticke, Pham-Dao, and Tjaden (2017) argue that the precautionary saving response to an increase in labor market risk causes households to substitute away from consumption expenditures into nonproductive, safe assets (such as government bonds), which can trigger a demand-driven recession.

These models differ in many important details, but they are all HANK models: they combine New Keynesian-style nominal rigidities with household heterogeneity and market incompleteness.

HANK: Central Elements

In the remainder of the paper, we focus on a version of HANK we developed with Benjamin Moll (Kaplan, Moll, and Violante 2018).[2] This formulation is distinctive in that it allows households to hold two assets: 1) a low-return liquid asset that represents holdings of cash, bank accounts, and government bonds, and 2) a high-return illiquid asset that is subject to a transaction cost and represents equities (which are mostly held in not-so-liquid retirement accounts), privately-owned businesses, and housing net worth. The household block of the model is based on Kaplan and Violante (2014). Households make decisions about labor supply, consumption, and savings. They face idiosyncratic labor productivity risk, which together with incomplete markets generate a precautionary saving motive.

Households can borrow in liquid assets up to an exogenous limit at an interest rate that is higher than the interest rate on liquid saving. We interpret this spread as an exogenous cost of financial intermediation. Inflows into liquid assets are after-tax labor earnings, interest payments on liquid assets, and lump-sum government transfers. Outflows from liquid assets are net deposits into the illiquid account, transaction costs, and consumption expenditures. Illiquid assets increase due to interest payments plus net deposits.

A trade-off between the two assets emerges endogenously. The low-return asset is ideal for consumption-smoothing (because of its liquidity properties), whereas the illiquid asset is preferred for long-term wealth accumulation (because of its high return).

The firm block of the model consists of a representative final-good producer that purchases a continuum of intermediate-goods in monopolistically competitive markets. The intermediate goods require capital and labor, which are rented from households in competitive input markets. Intermediate producers set prices to maximize their profits subject to convex costs of changing their price (as in Rotemberg 1982), which makes the price-level sticky. The illiquid asset held by households

[2] Here we provide only an intuitive description of the most important components of the model. In a companion working paper version, Kaplan and Violante (2018), we provide a more detailed description of the model, full details of computations presented in the following sections, and a number of additional analyses.

consists of both capital and shares that are claims to the equity of an aggregate portfolio of intermediate firms.[3]

The government raises revenue through a proportional tax on labor income. It uses the revenue to finance purchases of final goods, to pay lump-sum transfers to households, and to pay interest on its outstanding real debt. Through debt issuance, the government is the only provider of liquid assets in the economy. The monetary authority sets the nominal interest rate on liquid assets in accordance with a Taylor rule dictating that nominal rates rise when inflation rises, and fall when inflation falls.

The three equilibrium prices in this economy (the wage along with the returns on the liquid and illiquid assets) are determined by relevant market clearing conditions. In equilibrium, the return on illiquid assets is higher than the return on liquid assets in order to compensate households for the costs of transacting in the illiquid asset.

Several modeling choices that are inconsequential in RANK models can matter tremendously for the behavior of HANK models. In HANK, because of borrowing constraints and heterogeneity in marginal propensities to consume, both the timing and distribution of the fiscal transfers that are needed to balance the government budget constraint in the wake of a shock will matter. In RANK, because of Ricardian equivalence, the choice of fiscal rule does not matter. Similarly, the distribution of claims to firm profits, both across households and between liquid and illiquid assets, matters in HANK, whereas in RANK, profits are simply rebated to the representative household.[4] This also implies that in RANK models there is a unique stochastic discount factor for firms to use when setting prices, but in HANK models there is no unique discount factor. Also, in HANK, an assumption is needed about the extent to which fluctuations in labor demand are concentrated among different households, whereas in RANK no such assumption is necessary. Finally, because of the precautionary saving motive and occasionally binding borrowing constraint, in HANK the cyclicality of idiosyncratic uncertainty and access to liquidity are important determinants of the effects of aggregate shocks to household consumption (Acharya and Dogra 2018).

On the one hand, the sensitivity of HANK to these assumptions complicates the analysis and highlights important areas where micro data must be confronted. On the other hand, the assumptions about all these issues implicit in RANK models have little empirical support.[5]

Role of the Two Assets for Consumption Behavior

Virtually all of the existing HANK literature uses models with a single asset. However, we adopt the two-asset model because it is more successful at capturing key features of microeconomic consumption behavior.

[3] We assume that, within the illiquid account, resources can be freely moved between capital and equity, an assumption which allows us to reduce the dimensionality of the asset space.

[4] Broer, Hansen, Krussell, and Öberg (2016) discuss how the New Keynesian transmission mechanism is influenced by the assumptions that determine how profits get distributed across households.

[5] See the companion working paper, Kaplan and Violante (2018), for details on the specific assumptions we made in our baseline HANK model.

The coexistence of a low-return liquid asset and a high-return illiquid asset creates the conditions for the emergence of wealthy hand-to-mouth households (who hold little or no liquid wealth despite owning sizable amounts of illiquid assets) alongside poor hand-to-mouth households (who hold little net worth). The model is able to replicate the observation that around one-third of US households are hand-to-mouth with high marginal propensities to consume and, among these, around two-thirds are wealthy hand-to-mouth and one-third are poor hand-to-mouth (Kaplan, Violante, and Weidner 2014). The remaining households hold sufficient liquid wealth that their consumption dynamics are similar to those of the representative agent.

This existence of both types of hand-to-mouth households improves the fit of the model with respect to the responsiveness of consumption to interest rates and transitory income shocks. The two-asset model generates an average quarterly MPC out of small income windfalls of around 15 to 20 percent, as well as substantial heterogeneity in MPCs driven by heterogeneity in liquid wealth holdings across households. This level and distribution of MPCs is in line with the large body of evidence discussed earlier, as well as with more recent evidence in the context of the Great Recession (Mian, Rao, and Su 2013).

For comparison, the average MPC in an otherwise similar representative agent model is approximately equal to the discount rate, which is around 0.5 percent quarterly. When parameterized to match the same ratio of net worth to average income as in the data (and as in the two-asset model), the average quarterly MPC in the one-asset model is around 4 percent, which is eight times higher than in the representative agent model, but still much lower than empirical estimates.

Researchers have proposed modifications to the one-asset model to increase the average MPC to empirically realistic levels. One approach is to ignore all illiquid wealth and choose the household discount factor to generate the same ratio of *liquid wealth* to average income as in the data. Besides grossly misrepresenting observed household balance sheets, this approach also precludes the model from including capital—which is a crucial ingredient when analyzing macroeconomic dynamics in general equilibrium. A second approach used in (Carroll, Slacalek, Tokuoka, and White 2017; Krueger, Mitman, and Perri 2016) is to introduce enough heterogeneity in discount factors so that there are some very patient households that drive capital accumulation, together with some very impatient households that have a high MPC (although, even with heterogeneity in discount factors, a low-wealth calibration is usually required in order to generate a high aggregate MPC).

A problem with both these approaches is that, in order to generate realistic MPCs, the one-asset models feature many more *poor* hand-to-mouth households than are in the data. By abstracting from the illiquid assets held by the wealthy hand-to-mouth, these models also miss potentially important exposure of household consumption to fluctuations in returns to illiquid assets.

Comparison Between RANK and HANK

In this section, we compare the responses of representative agent and heterogeneous agent New Keynesian models to a series of aggregate shocks that are common in the study of business cycles. To allow for a clean comparison, we adopt a RANK model with the same two-asset structure, the same functional forms for preferences, technology, transaction costs, and price adjustment costs, and the same production side, government, and monetary authority as in HANK. The only important departure from HANK is the absence of any form of household heterogeneity.

We assume that each economy is initially in its steady state and is then hit by a one-time, unanticipated shock that is persistent and mean reverting. After the shock, the economies eventually return to their original steady states. Because the two models differ only on the household side, we focus our attention on the impulse response of aggregate consumption.

We start by analyzing three canonical sources of business cycles: demand, productivity, and monetary shocks. For consistency, we consider contractionary shocks whose size and persistence are chosen to generate a similar drop in aggregate consumption in the RANK model over the first quarter. For additional details of this comparison and the calibration of the two models, see Kaplan and Violante (2018).

Notions of Equivalence Between RANK and HANK

We define three notions of equivalence between RANK and HANK with respect to a given shock. The two models are *nonequivalent* when the impulse response function of consumption to a shock are different. They are *weakly equivalent* when the impulse responses are the same but the transmission mechanisms of the shock are different. They are *strongly equivalent* when both the impulse responses and the transmission mechanisms are the same. In other words, RANK and HANK are strongly equivalent in response to a given shock only if they produce the same impulse response function to the shock, for the same reasons.

Comparing impulse response functions across models, and hence identifying nonequivalence, is straightforward. Comparing transmission mechanisms, which is needed to distinguish between weak and strong equivalence, is open to some interpretation and various methods could be used. Here, we mostly emphasize a decomposition of the consumption impulse response function into the effects of all equilibrium objects that enter into the household consumption problem. These include wages, interest rates, asset prices, fiscal policy and the shock itself. A similar transmission mechanism requires this decomposition to be similar in the two models.

We also discuss two complementary approaches for comparing transmission mechanisms. First, we decompose the difference between the consumption responses in HANK and RANK into a general equilibrium discrepancy (due to different equilibrium price dynamics across models) and a partial equilibrium discrepancy (due to different sensitivity to the same price movements). A similar transmission mechanism requires both these discrepancies to be small in absolute value. Second, we compare the impulse response in HANK under alternative assumptions about the fiscal rule that balances the government budget constraint in the wake of the shock. In our baseline, changes in the stock of debt adjust to balance the budget in the short run and transfers adjust far in the future. Alternative fiscal rules imply different choices about the timing of the necessary adjustment in transfers. As explained earlier, in RANK, due to Ricardian equivalence, the choice of this fiscal rule has no effect on the impulse response function. Hence a similar transmission mechanism requires the timing of transfers to also have virtually no impact in HANK.

Demand Shocks: Strong Equivalence

Figure 1 compares the consumption response in HANK and RANK to a negative demand disturbance, modeled as a shock to households' marginal utility of consumption. Panel A shows that the impulse response functions for aggregate consumption are almost identical. In panels B and C, we plot the impulse response function decompositions for HANK and RANK, respectively. The decompositions are very similar in the two models, in the sense that by far the largest component of the decline in expenditures is the demand shock itself (the dash-dot line labeled "Pref"): expenditures fall because households become more patient and so postpone consumption. In Kaplan and Violante (2018), we show that the partial and general equilibrium discrepancies are both tiny, and that the aggregate consumption response is not affected by the fiscal rule.

Thus, the demand shock offers a clear-cut example of strong equivalence: both the aggregate response to the shock and its transmission mechanism are very similar in HANK and RANK.

Total Factor Productivity Shocks: Weak Equivalence

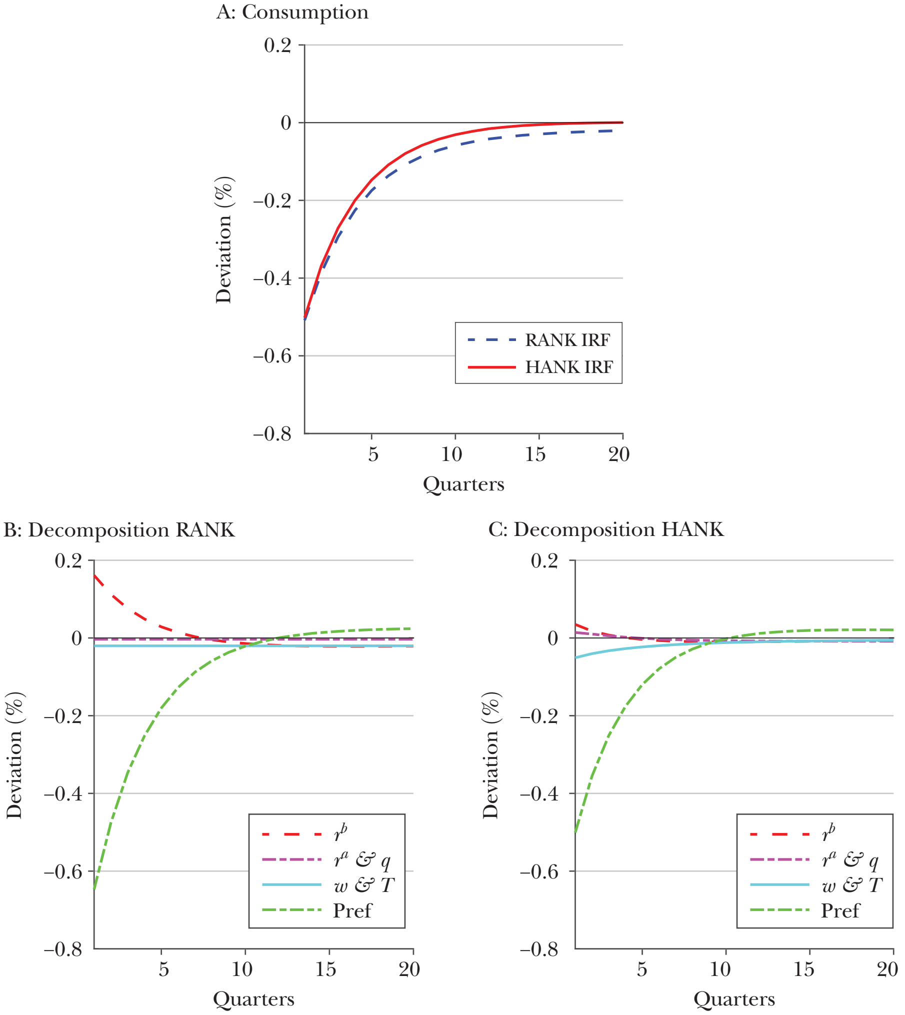
Figure 2 compares the consumption response in HANK and RANK to a negative technology shock, modeled as an unexpected drop in total factor productivity. As with the demand shock, panel A shows that the impulse response functions for the two models lie almost on top of each other.

However, here the transmission mechanisms are very different across models. The drop in productivity raises marginal costs and inflation, to which the central bank reacts by tightening monetary policy. The representative household responds to the higher interest rate by increasing liquid savings and postponing consumption. Thus, in RANK (panel B), the fall in consumption is driven entirely by intertemporal substitution in response to the higher interest rate. In HANK (panel C), the change in interest rates accounts for less than half of the fall in consumption. Instead, consumption falls mostly because disposable household income falls and the MPC out of a change in transitory income is large in HANK.[6] The productivity shock is thus an example of weak equivalence between HANK and RANK models. In

[6] As explained in Gali (1999), in RANK models, wages and hours rise in response to a contractionary shock to total factor productivity. This feature of New Keynesian models remains present in HANK. The fall in disposable household income accrues because of the fall in firm profits.

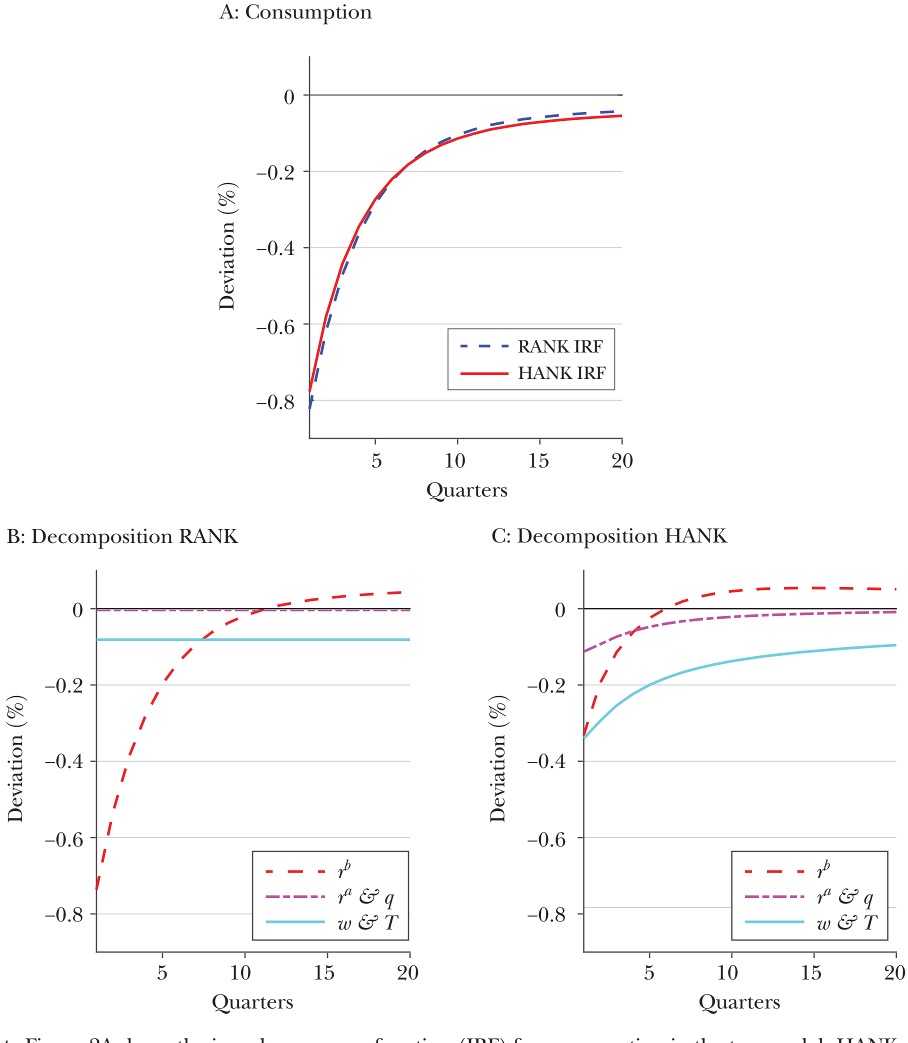
Figure 1

Negative Demand Shock in HANK and RANK: Impulse Response Functions (IRFs) for Consumption and their Decomposition



Note: Figure 1A shows the impulse response function for consumption in the two models HANK and RANK, while B and C present impulse response function decompositions. The line labeled “Pref” indicates the component of the impulse response due only to the preference shift, with all prices and transfers fixed at steady state values. The lines labeled r^b indicate the component of the impulse response due to the liquid rate changing, with all other prices, transfers, and the shock fixed at steady state values. Similarly, the lines labeled “r^a & q” indicate the component of the impulse response due to only the illiquid rate r^a and the equity price q changing, and the lines labeled “w & T” indicate the component of the impulse response due to only the wage w and lump-sum transfers T changing.

Figure 2
Negative Total Factor Productivity (TFP) Shock in HANK and RANK: Impulse Response Functions (IRFs) for Consumption and their Decomposition

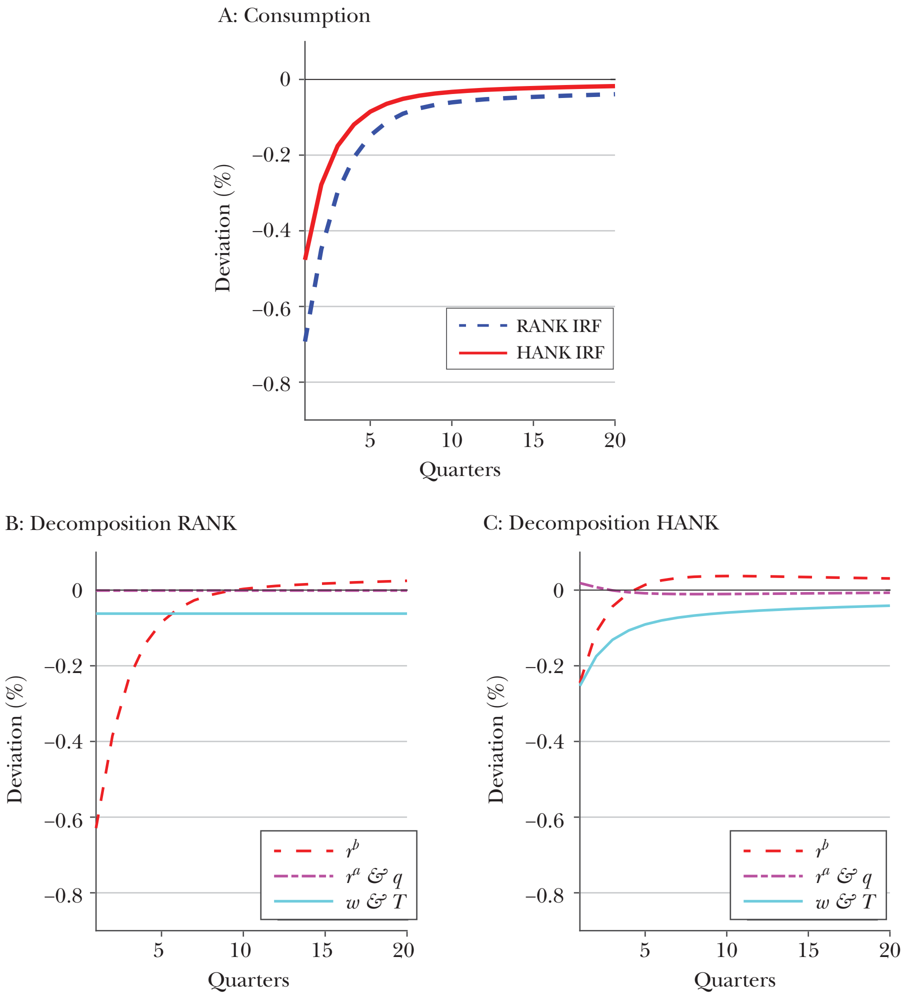


Note: Figure 2A shows the impulse response function (IRF) for consumption in the two models HANK and RANK, while B and C present impulse response function decompositions. The lines labeled r^b indicate the component of the impulse response due to the liquid rate changing, with all other prices and transfers fixed at steady state values. Similarly, the lines labeled "r^a & q" indicate the component of the impulse response due to only the illiquid rate r^a and the equity price q changing, and the lines labeled "w & T" indicate the component of the impulse response due to only the wage w and lump-sum transfers T changing.

Kaplan and Violante (2018), we show that both alternative approaches also suggest weak equivalence. In all three approaches, the differences in transmission mechanisms can be traced to the two hallmarks of the two-asset heterogeneous agent model that we discussed earlier: a high aggregate marginal propensity to consume out of income and a low sensitivity to interest rates.

Figure 3

Negative Monetary Shock (Positive Innovation to the Taylor Rule) in HANK and RANK: Impulse Response Functions (IRFs) for Consumption and their Decomposition



Note: Figure 3A shows the impulse response function for consumption in the two models HANK and RANK, while B and C present impulse response function (IRF) decompositions. The lines labeled r^b indicate the component of the impulse response due to the liquid rate changing (the shock), with all other prices and transfers fixed at steady state values. Similarly, the lines labeled "r^a & q" indicate the component of the impulse response due to only the illiquid rate r^a and the equity price q changing, and the lines labeled "w & T" indicate the component of the impulse response due to only the wage w and lump-sum transfers T changing.

Monetary Shock: Nonequivalence

Figure 3 compares the consumption response to a monetary policy shock in HANK and RANK, modeled as an innovation in the Taylor rule. Panel A shows that in the first quarter after the shock, consumption drops by roughly 50 percent more in RANK than in HANK. The transmission mechanism for monetary policy is different in the two models. In RANK (panel B), the direct intertemporal substitution channel due to the rise in the real liquid rate accounts for the whole effect. In HANK (panel C), the drop in consumption due to the fall in disposable income plays a role that is at least as important as the substitution channel. In Kaplan and Violante (2018), we spell out this difference in detail, and also show that the aggregate response is particularly sensitive to the choice of fiscal rule.[7] The monetary shock is thus an example of nonequivalence. Again, different sensitivities of household consumption to wages and interest rates are at the heart of the gap between the two impulse response functions.

Our result may appear to contrast with Werning (2015), who finds weak equivalence between the representative and heterogeneous agent model for the response of aggregate consumption to a monetary shock, but our findings are in fact consistent. His benchmark heterogeneous agent model is purposefully constructed so that the impulse response function for consumption following a change in the real rate is exactly the same as in RANK: the smaller partial equilibrium intertemporal substitution response to the change in interest rates in the heterogeneous agent model is exactly offset by the stronger aggregate demand response in general equilibrium. Werning illustrates how departures from his "as if" benchmark can lead to a larger or smaller aggregate consumption response to the monetary shock in HANK relative to RANK. Our version of HANK features several such departures, which explains why in our calibrated economy monetary shocks are examples of nonequivalence.

Fiscal Stimulus Shocks: Stark Nonequivalence

The large fiscal stimulus implemented by many governments in response to the Great Recession spurred a new wave of studies that made use of the emerging HANK framework (Oh and Reis 2012; McKay and Reis 2016; Hagadorn, Manovskii, and Mitman 2018). In this section, we show that fiscal stimulus is a stark example of nonequivalence between HANK and RANK models.

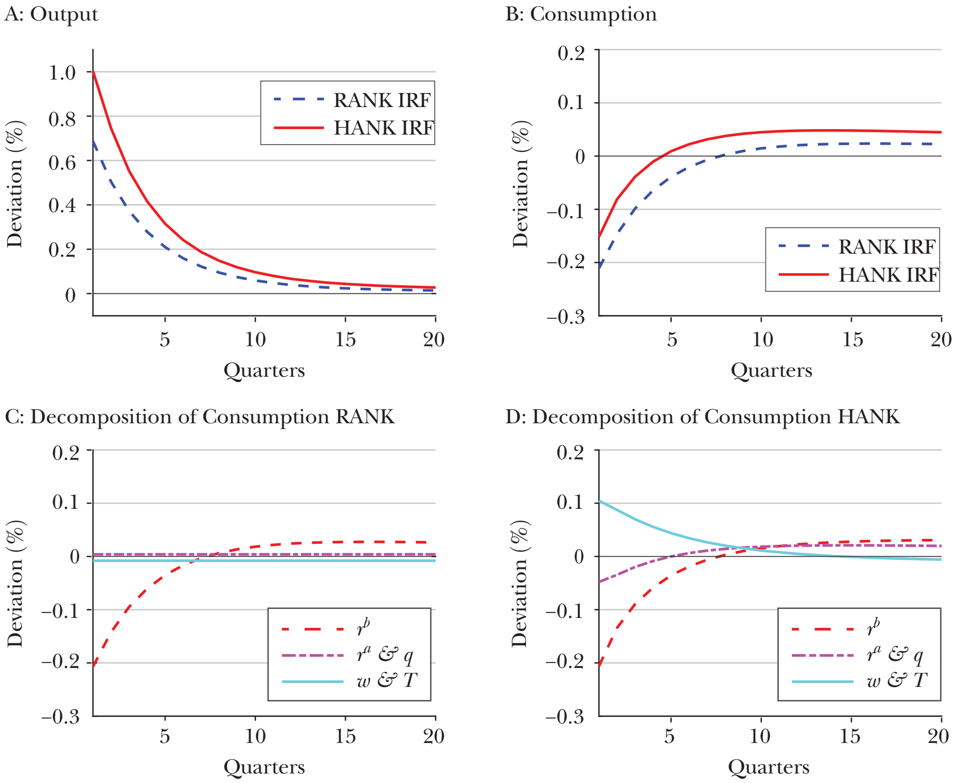
Figure 4 illustrates the effects of a deficit-financed temporary increase in government expenditures. Panel A shows that the expansionary effects on output are much stronger in HANK than in RANK, and panel B illustrates that the reason is the weaker crowding-out of private consumption. Crowding-out occurs because, in order to induce households to hold the additional debt issued by the government, the interest rate must rise. This puts downward pressure on private consumption.

The discrepancy between the two models in the transmission mechanism of the government expenditure shock can be seen in panels C and D. In RANK (panel C), the decline in aggregate consumption is entirely accounted for by the rise in the real interest rate (dashed line). In HANK (panel D), this decline is offset by the

[7] This result is especially stark for forward guidance shocks, as illustrated in Kaplan, Moll, and Violante (2016).

Figure 4

Fiscal Stimulus (Rise in Government Expenditures) in HANK and RANK: Impulse Response Functions (IRFs) for Output and Consumption and Decompositions for Consumption

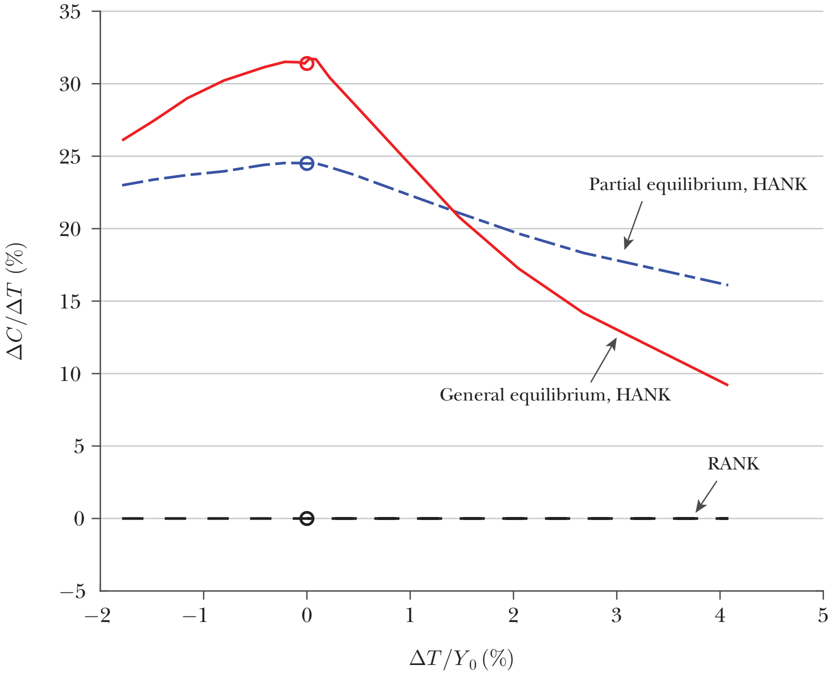


Note: Figure 4A and B show the impulse response functions (IRFs) for output and consumption in the two models HANK and RANK, while C and D present IRF decompositions for consumption. The lines labeled r^b indicate the component of the impulse response function due to the liquid rate changing, with all other prices and transfers fixed at the steady state values. Similarly, the lines labeled "r^a & q" indicate the component of the impulse response due to only the illiquid rate r^a and the equity price q changing, and the lines labeled "w & T" indicate the component of the impulse response due to only the wage w and lump-sum transfers T changing.

increase in labor demand and wages (solid line), which transmits strongly to household consumption through the high aggregate marginal propensity to consume.

Oh and Reis (2012) document that in the wake of the Great Recession, deficit-financed transfers were by far the largest component of fiscal stimulus in the United States. Figure 5 illustrates the effects of alternative temporary changes in lump-sum transfers (of different signs and sizes). The flat dashed line reminds us that, because of Ricardian neutrality, in RANK the consumption response is always zero. Thus, representative agent models are particularly ill suited for analyzing deficit-financed transfers. The dot-dash line shows the partial

Figure 5
Consumption Response to Change in Transfers in HANK and RANK



Note: The figure shows first quarter change in aggregate consumption (C) relative to first-quarter change in lump-sum transfers (T) in RANK and and in partial and general equilbirum HANK models. Y_0 is initial aggregate income.

equilibrium dynamics of aggregate consumption in the HANK model, which is simply the sum of the individual consumption responses, holding prices fixed at their steady-state levels. For expansionary transfer policies, the aggregate MPC falls with size because a larger fraction of the transfers is saved. For contractionary policies, the size-dependence is weaker (the line is flatter to the left of zero) since smoothing the fall in income for many households would now require tapping into expensive credit. These predictions are in line with the evidence discussed earlier both qualitatively (in terms of size-dependence and sign asymmetries) and quantitatively (the quarterly aggregate MPC is around 20 percent).

The solid line illustrates that in the full HANK model, for a wide range of values, the general equilibrium response is stronger due to the aggregate demand effects. However, since the model features an active Taylor rule, a very large stimulus can be so inflationary that the monetary authority raises interest rates to a point that it overcompensates for the expansionary effects of fiscal policy.

Simpler Models that Mimic HANK

We have repeatedly seen that the key differences between HANK and RANK models that lead to nonequivalence or weak equivalence can be traced back to the lower sensitivity of consumption to interest rates and higher sensitivity to disposable income. A natural question that arises is whether some simple modifications to RANK could replicate these features of consumption behavior and thus generate transmission mechanisms that are similar to those in HANK without the computational complexity of a full-blown heterogeneous agent model.

One such modification is the Two-Agent New Keynesian model (TANK), based on the spender–saver model of Campbell and Mankiw (1989). Early examples of this approach are Iacoviello (2005), Galí, López-Salido, and Vallés (2007), and Bilbiie (2008). For certain shocks, TANK can approach strong equivalence with HANK and thus offer a useful shortcut. For other questions, such as the macroeconomic impact of fiscal transfers of different sizes and signs, the two models yield different answers. In Kaplan, Moll, and Violante (2018) and Bilbiie (2017), similarities between HANK and TANK are discussed in the context of monetary policy shocks, and Debortoli and Galí (2017) extend the comparison to various other shocks and fiscal rules.

An alternative avenue for modifying RANK is to introduce liquid wealth directly into the utility function of the representative household. This shortcut captures, in a reduced-form way, the idea that in the presence of uninsurable risk, the household sector as a whole values the existence of a supply of safe, liquid assets because of its precautionary value (as in Aiyagari and McGrattan 1998). In Kaplan and Violante (2018), we show that this augmented RANK model has several other properties that bring it closer to the HANK model (see also Michaillat and Saez 2018).

Macro Questions that Require a Model with Heterogeneity

So far, we have addressed macroeconomic questions about impulse and propagation that are well-posed in both heterogeneous and representative agent models. However, some questions pertaining to macroeconomic dynamics can only be addressed in models with household heterogeneity. In this section, we provide three examples: the effects of aggregate shocks that are not well-defined in representative agent models; how different responses to aggregate shocks by households at different parts of the distribution can aid in the identification of shocks and transmission mechanisms; and the effect of aggregate shocks on household inequality.[8]

Microfoundations of a Fall in Aggregate Demand

Two salient features of the Great Recession were a deep and prolonged drop in expenditures and a sharp fall in the nominal interest rate that led to a binding zero lower bound. These features of the data are consistent with a drop in aggregate demand

[8] In Kaplan and Violante (2018), we provide more details on all these exercises and some additional figures.

as a primary driving force behind the recession. To generate a large sudden fall in aggregate demand in representative agent models, most researchers have resorted to assuming a shock to the discount factor of the representative household. This type of shock was the basis of the earlier discussion summarized in Figure 1. Macroeconomists often justify this shock as a stand-in for some unspecified deeper force that acts as if "households become more patient" (Eggertsson and Krugman 2012).

HANK models offer the possibility to generate a rise in households' desire to save through mechanisms that are both more micro-founded and consistent with aspects of micro data. One leading example is tighter credit limits that reduce borrowing capacity, leading constrained households to deleverage sharply and leading unconstrained households to increase their savings in order to avoid being constrained in the future (as in Guerrieri and Lorenzoni 2017). Another example is a surge in uninsurable labor market risk, which exacerbates the precautionary saving motive (as in Den Haan, Rendahl, and Riegler 2017; Bayer et al. 2017). In the presence of sticky prices, both types of shocks induce a fall in aggregate expenditures and a large enough drop in the real interest rate that the zero lower bound on nominal rates binds.[9]

For both of these representations of a shortfall in aggregate demand, the two-asset version of HANK offers an important advantage over its one-asset counterpart. In the aftermath of the shock, the additional household savings are channeled towards the unproductive liquid asset, which is the better asset for consumption smoothing purposes, rather than towards productive illiquid capital, thus avoiding a counterfactual investment boom. Indeed, the literature that studies these shocks in one-asset HANK models typically abstracts from capital for this reason.

Heterogeneity in the Transmission Mechanism

As explained earlier, models can differ in their transmission mechanism while not differing in terms of their aggregate response to certain shocks. Hence, collecting empirical evidence on the mechanism itself is crucial in distinguishing between models. Time-series data alone might not be that useful because confounding factors abound. An alternative approach is to use cross-sectional data (as discussed by Nakamura and Steinsson in this issue). In this context, one advantage of heterogeneous agent models is that they make predictions about how the effect of an aggregate shock varies across the distribution of households. One can therefore exploit rich micro data to gather support for a specific model or mechanism.

For example, in our two-asset version of HANK, the consumption drop in response to a contractionary monetary shock differs tremendously across households depending on their holdings of liquid wealth. For the mass of hand-to-mouth households with zero liquid wealth, the response is largest and is almost entirely due to the general equilibrium drop in their labor income. But for households

[9] In certain models that admit aggregation in closed form, it is possible to show that a rise in idiosyncratic uncertainty is formally equivalent to a rise in the discount factor of the pseudo-representative agent (Braun 2012).

with substantial positive liquid wealth, the direct effect of the interest rate change is larger than the effect of the drop in their labor income, because these consumers have a low marginal propensity to consume but a high sensitivity to interest rate changes. Empirical work using household panel data on consumption, income, and wealth provides some support for this pattern of cross-sectional transmission mechanism (Cloyne, Ferreira, and Surico 2016).

Examining the consumption response of aggregate shocks at different points in the distribution of households is also a promising avenue to identify the underlying sources of aggregate fluctuations. For example, the three types of aggregate demand disturbances just described—preferences, credit tightness, and income risk—all produce qualitatively similar aggregate dynamics: a large reduction in aggregate expenditures that leads to a decline in interest rates. However, the distributional response of these three shocks is very different: The discount factor shock generates consumption responses that are much more evenly distributed across the liquid wealth distribution than either the credit or risk shocks. And relative to the risk shock, the credit shock generates a consumption response that is more heavily concentrated among households with negative liquid wealth. In Kaplan and Violante (2018), we illustrate these differences.

Impact of Aggregate Shocks on Inequality

Heterogeneous agent models are not only valuable for understanding how wealth and income inequality can affect the magnitude and transmission mechanism of aggregate shocks. They are also useful when the question is turned on its head: to what extent do macroeconomic shocks affect inequality?

For example, consider the effects of a contractionary monetary shock on the distribution of consumption in the two-asset HANK model. The rise in the interest rate pushes up consumption of the very wealthy households through a positive income effect. The equilibrium fall in aggregate demand leads to a reduction in labor income, which lowers consumption most sharply for households at the bottom of the distribution. In Kaplan and Violante (2018), we illustrate the quantitative strength of these forces and conclude that the monetary shock has only a modest effect on consumption dispersion that persists as long as the shock itself does. The empirical analysis in Coibion, Gorodnichenko, Kueng, and Silvia (2017) finds some support for this finding that contractionary monetary policy has a positive, but small, impact on inequality.

Conclusions: Looking Ahead

A new macroeconomic framework is emerging. It embeds a rich representation of household consumption and portfolio choices, consistent with many aspects of microeconomic data, into a dynamic general equilibrium model of the macroeconomy that can accommodate a wide range of aggregate shocks and demand-side effects. This framework offers a coherent way to study questions that pertain to cross-sectional inequality, economic mobility, social insurance, and redistributive policies as well as traditional business cycle questions that bear on the dynamics of macroeconomic variables, propagation mechanisms of aggregate shocks, and stabilization policies.

This framework is still in its infancy. To conclude this essay, we outline several promising directions for the development of this class of models.

New Keynesian models rely on wage and price stickiness to explain both why monetary policy can have real effects and why aggregate demand can affect real output. A promising alternative aggregate demand channel, which does not rely on price stickiness, is based on search frictions in the product market. Households vary the effort with which they hunt for bargains depending on their wealth, income, and demand for consumption. Heterogeneous agent models with search in product markets can embed this mechanism and generate aggregate demand effects either through endogenous movements in the competitiveness of product markets and markups (Kaplan and Menzio 2016) or through endogenous movements in aggregate productivity (Huo and Ríos-Rull, 2016).

In existing HANK models, labor market risk is mostly exogenous. Labor market frictions are one way to provide micro foundations for the extent and nature of idiosyncratic labor market risk. For example, Hubmer (2018) shows that skewness in earnings growth uncovered in micro data (Guvenen, Schulhofer-Wohl, Song, and Yogo 2015; Arellano, Blundell, and Bonhomme 2017) arises endogenously in a canonical frictional model of the labor market with on-the-job search. As another example, Moscarini and Postel-Vinay (2017) describe a setting where firms choose to match outside offers to retain their workers, in which case, the wage goes up without any change in productivity, generating inflation. Embedding this mechanism into a heterogeneous agent model could then generate a credible micro-foundation for the two main driving forces behind inflation dynamics: 1) aggregate demand shocks driven by the distribution of marginal propensities to consume and 2) cost-push shocks driven by the distribution of workers along the job ladder.

The HANK model analyzed in this essay is a model of net household asset positions rather than gross positions. In reality, many households hold highly leveraged portfolios, particularly with regards to illiquid assets, such as housing. If mortgage contracts allow for some degree of pass-through of interest rates (either because of adjustable rates or the option to refinance), then changes in interest rates can have significant cash-flow effects on expenditures for borrowers (for example, Flodén et al. 2016; Di Maggio et al. 2017; La Cava, Hughson, and Kaplan 2016). In addition, many assets and liabilities (like cash, bank accounts, government bonds, secured and unsecured debt) earn nominal returns that do not adjust instantaneously to aggregate conditions, and so surprise inflation can have redistributive effects (Doepke and Schneider 2006; Auclert 2017). And when households have long-term nominal debt contracts (as is the typically the case for mortgages), then anticipated inflation can also have redistributive effects. In a version of the model with endogenous credit limits, aggregate shocks would transmit to the real economy also by modifying the extent of credit availability (for example, Chatterjee, Corbae, Nakajima, and Ríos-Rull

2007; Agarwal, Chomsisengphet, Mahoney, and Stroebel 2015; Gross, Notowidigdo, and Nakajima 2016; Favilukis, Ludvigson, and Van Nieuwerburgh 2017).

In this first generation of HANK models, equity prices barely move in response to aggregate shocks (for example, monetary shocks), and when they do, it is often in the wrong direction. The most promising way to generate realistic asset price movements in response to macroeconomic fluctuations is through time-varying risk premia: that is, the willingness of market participants to bear risk is greater in booms than in recessions (Cochrane 2017). Future versions of these models should aim to generate large and variable risk premia, as well as to recognize that some households are much more exposed to asset price movements than are others because of the composition of their balance sheets (Mian, Rao, and Su 2013; Glover, Heathcote, Krueger, and Ríos-Rull 2017) and the nature of their labor income (Guvenen, Karahan, Ozkan, and Song 2015).

There are no banks in the baseline HANK models: liquid assets are provided directly by the fiscal authority and backed by future tax revenues. Any changes in households' demand to save in liquid assets therefore directly affect the government budget constraint, which induces a stronger link between fiscal policy and household savings behavior than in reality. Moreover, many of the prevailing accounts of the Great Recession attribute a central role to the deterioration of banks' balance sheets. Exploring this latter propagation mechanism requires an explicit model of the banking sector, along with regulatory constraints on bank balance sheets. The two-asset version of HANK lends itself naturally to the introduction of banks, since one of the key roles of financial intermediaries is transformation of assets from higher to lower liquidity (as illustrated in Kaplan, Moll, and Violante 2016).

A heterogeneous agent model could help to explore deviations from rational expectations and complete information. Some recent papers have showed how dispersed information (Angeletos and Lian 2017) or behavioral biases (Farhi and Werning 2017) can have consequences for the relative strengths of partial equilibrium versus general equilibrium effects of aggregate shocks, thus changing the incidence of shocks across the household distribution.

Finally, the analysis of optimal policy changes drastically in a heterogeneous agent economy because redistributive and social insurance implications come into play. For example, McKay and Reis (2016) show that removing automatic fiscal stabilizers would not amplify aggregate consumption fluctuations as long as monetary policy follows a standard Taylor rule, but could lead to large welfare costs because of the decrease in social insurance. Gornemann, Kuester, and Nakajima (2016) argue that a monetary policy rule that emphasizes price stability redistributes towards rich households, while one that stresses output stability redistributes towards poor households who are more exposed to unemployment risk, and that the median household prefers output stability. An emerging literature is making progress towards characterizing optimal policies in this class of models (for example, Le Grand and Ragot 2017; Nuño and Thomas 2016; Bhandari, Evans, Golosov, and Sargent 2017).

■ *We are grateful to Felipe Alves for outstanding research assistance and to Mark Gertler, David Lopez-Salido, Virgiliu Midrigan, Ben Moll, Matt Rognlie, Tom Sargent, and Christian Wolf for comments.*

References

Acharya, Sushant, and Keshav Dogra. 2018. "Understanding HANK: Insights from a PRANK." Federal Reserve Bank of New York Staff Report 835. February.

Agarwal, Sumit, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel. 2015. "Regulating Consumer Financial Products: Evidence from Credit Cards." *Quarterly Journal of Economics* 130(1): 111–64.

Ahn, SeHyoun, Greg Kaplan, Benjamin Moll, Thomas Winberry, and Christian Wolf. 2017. "When Inequality Matters for Macro and Macro Matters for Inequality." Chap. 1 in *NBER Macroeconomics Annual 2017*, vol. 32. University of Chicago Press.

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

Aiyagari, S. Rao, and Ellen R. McGrattan. 1998. "The Optimum Quantity of Debt." *Journal of Monetary Economics* 42(3): 447–69.

Angeletos, George-Marios, and Chen Lian. 2017. "Dampening General Equilibrium: From Micro to Macro." NBER Working Paper 23379.

Arellano, Manuel, Richard Blundell, and Stéphane Bonhomme. 2017. "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework." *Econometrica* 85(3): 693–734.

Attanasio, Orazio, and Steven J. Davis. 1996. "Relative Wage Movements and the Distribution of Consumption." *Journal of Political Economy* 104(6): 1227–62.

Auclert, Adrien. 2017. "Monetary Policy and the Redistribution Channel." NBER Working Paper 23451.

Baker, Scott. Forthcoming. "Debt and the Response to Household Income Shocks: Validation and Application of Linked Financial Account Data." *Journal of Political Economy*.

Bayer, Christian, Ralph Lütticke, Lien Pham-Dao, and Volker Tjaden. 2017. "Precautionary Savings, Illiquid Assets, and the Aggregate Consequences of Shocks to Household Income Risk." Working Paper, University of Bonn.

Benhabib, Jess, and Alberto Bisin. Forthcoming. "Skewed Wealth Distributions: Theory and Empirics." *Journal of Economic Literature*.

Bewley, Truman. 1983. "A Difficulty with the Optimum Quantity of Money." *Econometrica* 51(5): 1485–1504.

Bhandari, Anmol, David Evans, Mikhail Golosov, and Thomas J. Sargent. 2017. "Inequality, Business Cycles and Monetary-Fiscal." Working Paper, University of Minnesota.

Bilbiie, Florin O. 2008. "Limited Asset Markets Participation, Monetary Policy and (Inverted) Aggregate Demand Logic." *Journal of Economic Theory* 140(1): 162–96.

Bilbiie, Florin Ovidiu. 2017. "The New Keynesian Cross: Understanding Monetary Policy with Hand-to-Mouth Households." CEPR Discussion Paper no. DP11989. Available at SSRN: https://ssrn.com/abstract=2957528.

Blundell, Richard, Luigi Pistaferri, and Ian Preston. 2008. "Consumption Inequality and Partial Insurance." *American Economic Review* 98(5): 1887–1921.

Boppart, Timo, Per Krusell, and Kurt Mitman. 2017. "Exploiting MIT Shocks in Heterogeneous-Agent Economies: The Impulse Response as a Numerical Derivative." NBER Working Paper 24138.

Braun, Anton, and Tomoyuki Nakajima. 2012. "Uninsured Countercyclical Risk: An Aggregation Result and Application to Optimal Monetary Policy." *Journal of the European Economic Association* 10(6): 1450–74.

Broda, Christian, and Jonathan A. Parker. 2014. "The Economic Stimulus Payments of 2008 and the Aggregate Demand for Consumption." *Journal of Monetary Economics* 68(S): 20–36.

Broer, Tobias, Niels-Jakob H. Hansen, Per Krusell, and Erik Öberg. 2016. "The New Keynesian Transmission Mechanism:

A Heterogeneous-Agent Perspective." NBER Working Paper 22418.

Browning, Martin, and Thomas F. Crossley. 2001. "The Life-Cycle Model of Consumption and Saving." *Journal of Economic Perspectives* 15(3): 3–22.

Brumm, Johannes, and Simon Scheidegger. 2017. "Using Adaptive Sparse Grids to Solve High-Dimensional Dynamic Models." *Econometrica* 85(5): 1575–1612.

Caballero, Ricardo J. 1999. "Aggregate Investment." Chap. 12 in *Handbook of Macroeconomics*, vol. 1B," edited by John B. Taylor and Michael Woodford. North Holland.

Campbell, John Y., and N. Gregory Mankiw. 1989. "Consumption, Income and Interest Rates: Reinterpreting the Time Series Evidence." In *NBER Macroeconomics Annual,* vol. 4, 185–246, edited by Olivier Jean Blanchard and Stanley Fischer. MIT Press.

Canzoneri, Matthew B., Robert E. Cumby, and Behzad T. Diba. 2007. "Euler Equations and Money Market Interest Rates: A Challenge for Monetary Policy Models." *Journal of Monetary Economics* 54(7): 1863 –81.

Carroll, Christopher D. 1997. "Buffer-Stock Saving and the Life Cycle/Permanent Income Hypothesis." *Quarterly Journal of Economics* 112(1): 1–55.

Carroll, Christopher, Jiri Slacalek, Kiichi Tokuoka, and Matthew N. White. 2017. "The Distribution of Wealth and the Marginal Propensity to Consume." *Quantitative Economics* 8(3): 977–1020.

Chatterjee, Satyajit, Dean Corbae, Makoto Nakajima, and José-Víctor Ríos-Rull. 2007. "A Quantitative Theory of Unsecured Consumer Credit with Risk of Default." *Econometrica* 75(6): 1525–89.

Christelis, Dimitris, Dimitris Georgarakos, Tullio Jappelli, Luigi Pistaferri, and Maarten van Rooij. 2017. "Asymmetric Consumption Effects of Transitory Income Shocks." Working Paper 467, Centre for Studies in Economics and Finance (CSEF).

Clarida, Richard, Jordi Gali, and Mark Gertler. 1999."The Science of Monetary Policy: A New Keynesian Perspective." *Journal of Economic Literature* 37(4): 1661–1707.

Cloyne, James, Clodomiro Ferreira, and Paolo Surico. 2016. "Monetary Policy When Households Have Debt: New Evidence on the Transmission Mechanism." Bank of England Working Paper 589. Available at SSRN https://ssrn.com/abstract=2765415.

Cochrane, John H. 1991. "A Simple Test of Consumption Insurance." *Journal of Political Economy* 99(5): 957–76.

Cochrane, John. 2015. "Whither Inflation." *Grumpy Economist* blog, August 31. https://johnhcochrane.blogspot.com/2015/08/whither-inflation.html.

Cochrane, John H. 2017. "Macro-Finance." *Review of Finance* 21(3): 945–85.

Coibion, Olivier, Yuriy Gorodnichenko, Lorenz Kueng, and John Silvia. 2017. "Innocent Bystanders? Monetary Policy and Inequality." *Journal of Monetary Economics* 88: 70–89.

Costancio, Vitor. 2017. "Inequality and Macroeconomic Policies." Speech at the Annual Congress of the European Economic Association, Lisbon, Portugal, August 22. https://www.ecb.europa.eu/press/key/date/2017/html/ecb.sp170822.en.html.

Deaton, Angus. 1987. "Life-cycle Models of Consumption: Is the Evidence Consistent with the Theory?" Chap. 14 in *Advances in Econometrics:* Vol. 2: *Fifth World Congress,* edited by Truman F. Bewley. Cambridge University Press.

Debortoli, Davide, and Jordi Galí. 2017. "Monetary Policy with Heterogeneous Agents: Insights from TANK Models." Technical Report.

De Nardi, Mariacristina, and Giulio Fella. 2017. "Saving and Wealth Inequality." *Review of Economic Dynamics* 26: 280–300.

Den Haan, Wouter J., Pontus Rendahl, and Markus Riegler. 2017. "Unemployment (Fears) and Deflationary Spirals." *Journal of the European Economic Association,* advance article, https://doi.org/10.1093/jeea/jvx040.

Di Maggio, Marco, Amir Kermani, Benjamin J. Keys, Tomasz Piskorski, Rodney Ramcharan, Amit Seru, and Vincent Yao. 2017. "Interest Rate Pass-Through: Mortgage Rates, Household Consumption, and Voluntary Deleveraging." *American Economic Review* 107(11): 3550–88.

Doepke, Matthias, and Martin Schneider. 2006. "Inflation and the Redistribution of Nominal Wealth." *Journal of Political Economy* 114(6): 1069–97.

Duarte, Victor. 2018. "Machine Learning for Continuous-Time Economics." Available at SSRN: https://ssrn.com/abstract=3012602.

Eggertsson, Gauti B., and Paul Krugman. 2012. "Debt, Deleveraging, and the Liquidity Trap: A Fisher–Minsky–Koo Approach." *Quarterly Journal of Economics* 127(3): 1469–1513.

Fagereng, Andreas, Martin Holm, and Gisle Natvik. 2016. "MPC Heterogeneity and Household Balance Sheets." Discussion Paper 852, Statistics Norway.

Farhi, Emmanuel, and Iván Werning. 2017. "Monetary Policy, Bounded Rationality, and Incomplete Markets." NBER Working Paper 23281.

Favilukis, Jack, Sydney C. Ludvigson, and Stijn

Van Nieuwerburgh. 2017. "The Macroeconomic Effects of Housing Wealth, Housing Finance, and Limited Risk Sharing in General Equilibrium." *Journal of Political Economy* 125(1): 140–223.

Fernández-Villaverde, Jesús, Samuel Hurtado, and Galo Nuño. 2018. "Financial Frictions and the Wealth Distribution." https://www.sas.upenn.edu/~jesusfv/Financial_Frictions_Wealth_Distribution.pdf.

Flodén, Martin, Matilda Kilström, Jósef Sigurdsson, and Roine Vestman. 2016. "Household Debt and Monetary Policy: Revealing the Cash-Flow Channel." Swedish House of Finance Research Paper no. 16-8. Available at SSRN: https://ssrn.com/abstract=2748232.

Fuster, Andreas, Greg Kaplan, and Basit Zafar. 2018. "What Would You Do With $500? Evidence from Gains, Losses, News and Loans." Unpublished paper, University of Chicago.

Galí, Jordi. 1999. "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?" *American Economic Review* 89(1): 249–71.

Galí, Jordi, J. David López-Salido, and Javier Vallés. 2007. "Understanding the Effects of Government Spending on Consumption." *Journal of the European Economic Association* 5(1): 227–70.

Gertler, Mark, and Simon Gilchrist. 1994. "Monetary Policy, Business Cycles, and the Behavior of Small Manufacturing Firms." *Quarterly Journal of Economics* 109(2): 309–340.

Glover, Andrew, Jonathan Heathcote, Dirk Krueger, and José-Víctor Ríos-Rull. 2017. "Intergenerational Redistribution in the Great Recession." Working Paper, University of Texas. https://www.sas.upenn.edu/~dkrueger/research/RecessionNew.pdf.

Gornemann, Nils, Keith Kuester, and Makoto Nakajima. 2016. "Doves for the Rich, Hawks for the Poor? Distributional Consequences of Monetary Policy." International Finance Discussion Papers 1167. https://www.federalreserve.gov/econresdata/ifdp/2016/files/ifdp1167.pdf.

Gross, Tal, Matthew J. Notowidigdo, and Jialan Wang. 2016. "The Marginal Propensity to Consume over the Business Cycle." NBER Working Paper 22518.

Guerrieri, Veronica, and Guido Lorenzoni. 2017. "Credit Crises, Precautionary Savings, and the Liquidity Trap." *Quarterly Journal of Economics* 132(3): 1427–67.

Guvenen, Fatih. 2011. "Macroeconomics with Heterogeneity: A Practical Guide." *Economic Quarterly, Federal Reserve Bank of Richmond* 97(3): 255–326.

Guvenen, Fatih, Fatih Karahan, Serdar Ozkan, and Jae Song. 2015. "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Risk?" NBER Working Paper 20913.

Guvenen, Fatih, Sam Schulhofer-Wohl, Jae Song, and Motohiro Yogo. 2017. "Worker Betas: Five Facts about Systematic Earnings Risk." *American Economic Review* 107(5): 398–403.

Hagedorn, Marcus, Iourii Manovskii, and Kurt Mitman. 2018. "The Fiscal Multiplier." Working Paper, University of Oslo. https://www.sas.upenn.edu/~manovski/papers/The_Fiscal_Multiplier.pdf.

Hall, Robert. 1978. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *Journal of Political Economy* 86(6): 971–987.

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. 2009. "Quantitative Macroeconomics with Heterogeneous Households." *Annual Review of Economics* 1: 319–54.

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. 2014. "Consumption and Labor Supply with Partial Insurance: An Analytical Framework." *American Economic Review* 104(7): 2075–2126.

Hubmer, Joachim. 2018. "The Job Ladder and its Implications for Earnings Risk." *Review of Economic Dynamics* 29: 172–94.

Huggett, Mark. 1993. "The Risk-Free Rate in Heterogeneous-Agent Incomplete-Insurance Economies." *Journal of Economic Dynamics and Control* 17(5–6): 953–69.

Huo, Zhen, and José Víctor Ríos-Rull. 2016. "Financial Frictions, Asset Prices, and the Great Recession." CEPR Discussion Paper DP11544. Available at SSRN: https://ssrn.com/abstract=2847078.

Iacoviello, Matteo. 2005. "House Prices, Borrowing Constraints, and Monetary Policy in the Business Cycle." *American Economic Review* 95(3): 739–64.

Imrohoroğlu, Ayşe. 1989. "Cost of Business Cycles with Indivisibilities and Liquidity Constraints." *Journal of Political Economy* 97(6): 1364–83.

Jappelli, Tullio, and Luigi Pistaferri. 2010. "The Consumption Response to Income Changes." *Annual Review of Economics* 2: 479–506.

Johnson, David S., Jonathan A. Parker, and Nicholas S. Souleles. 2006. "Household Expenditure and the Income Tax Rebates of 2001." *American Economic Review* 96(5): 1589–1610.

Kaplan, Greg, and Guido Menzio. 2016. "Shopping Externalities and Self-Fullfilling Unemployment Fluctuations." *Journal of Political Economy* 124(3): 771–825.

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante. 2016. "A Note on Unconventional

Monetary Policy in HANK." Technical Report, Princeton University.

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante. 2018. "Monetary Policy According to HANK." *American Economic Review* 108(3): 697–743.

Kaplan, Greg, and Giovanni L. Violante. 2014. "A Model of the Consumption Response to Fiscal Stimulus Payments." *Econometrica* 82(4): 1199–1239.

Kaplan, Greg, and Gianluca L. Violante. 2018. "Microeconomic Heterogeneity and Macroeconomic Shocks." NBER Working Paper 24734.

Kaplan, Greg, Gianluca Violante, and Justin Weidner. 2014. "The Wealthy Hand-to-Mouth." *Brookings Papers on Economic Activity*, no. 1, pp. 77–138.

Khan, Aubhik, and Julia K. Thomas. 2008. "Idiosyncratic Shocks and the Role of Nonconvexities in Plant and Aggregate Investment Dynamics." *Econometrica* 76(2): 395–436.

Krueger, Dirk, Kurt Mitman, and Fabrizio Perri. 2016. "Macroeconomics and Household Heterogeneity." Chap. 11 in *Handbook of Macroeconomics*, vol. 2A, edited by John B. Taylor and Harald Uhlig. Elsevier.

Krusell, Per, and Anthony A. Smith. 1998. "Income and Wealth Heterogeneity in the Macroeconomy." *Journal of Political Economy* 106(5): 867–96.

Kuroda, Haruiko. 2017. "Opening Remarks at the 2017 BOJ–IMES Conference Hosted by the Institute for Monetary and Economic Studies, Bank of Japan." May 24. https://www.boj.or.jp/en/announcements/press/koen_2017/ko170524a.htm.

La Cava, Gianni, Helen Hughson, and Greg Kaplan. 2016. "The Household Cash Flow Channel of Monetary Policy." Research Discussion Paper 2016-12, Reserve Bank of Australia. http://www.rba.gov.au/publications/rdp/2016/pdf/rdp2016-12.pdf.

Le Grand, François, and Xavier Ragot. 2017. "Optimal Fiscal Policy with Heterogeneous Agents and Aggregate Shocks." 2017 Meeting Paper 969, Society for Economic Dynamics.

Ljungqvist, Lars, and Thomas J. Sargent. 2004. *Recursive Macroeconomic Theory*, MIT Press Books.

McKay, Alisdair, Emi Nakamura, and Jón Steinsson. 2016. "The Power of Forward Guidance Revisited." *American Economic Review* 106(10): 3133–58.

McKay, Alisdair, and Ricardo Reis. 2016. "The Role of Automatic Stabilizers in the U.S. Business Cycle." *Econometrica* 84(1): 141–94.

Mian, Atif, Kamalesh Rao, and Amir Su. 2013. "Household Balance Sheets, Consumption, and the Economic Slump." *Quarterly Journal of Economics* 128(4): 1687–1726.

Michaillat, Pascal, and Emmanuel Saez. 2018. "A New Keynesian Model with Wealth in the Utility Function." https://eml.berkeley.edu/~saez/michaillat-saezFeb18WUNK.pdf.

Misra, Kanishka, and Paolo Surico. 2014. "Consumption, Income Changes, and Heterogeneity: Evidence from Two Fiscal Stimulus Programs." *American Economic Journal: Macroeconomics* 6(4): 84–106.

Moscarini, Giuseppe, and Fabien Postel-Vinay. 2017. "The Job Ladder: Inflation vs. Reallocation." Working Paper, Yale University.

Nuño, Galo, and Carlos Thomas. 2016. "Optimal Monetary Policy with Heterogeneous Agents." Working Paper 1624, Bank of Spain.

Oh, Hyunseung, and Ricardo Reis. 2012. "Targeted Transfers and the Fiscal Response to the Great Recession." *Journal of Monetary Economics* 59(S): 50–64.

Ottonello, Pablo, and Thomas Winberry. 2018. "Financial Heterogeneity and the Investment Channel of Monetary Policy." NBER Working Paper 24221.

Pröhl, Elisabeth. 2017. "Approximating Equilibria with Ex-Post Heterogeneity and Aggregate Risk." Swiss Finance Institute Research Paper 17-63. Available at SSRN: https://ssrn.com/abstract=2620937.

Quadrini, Vincenzo, and José-Víctor Ríos-Rull. 2015. "Inequality in Macroeconomics." Chap. 14 in *Handbook of Income Distribution*, vol. 2B, edited by Anthony B. Atkinson and François Bourguignon, pp. 1229–1302. Elsevier.

Reiter, Michael. 2009. "Solving Heterogeneous-Agent Models by Projection and Perturbation." *Journal of Economic Dynamics and Control* 33(3): 649–65.

Ríos-Rull, José-Víctor. 1995. "Models with Heterogeneous Agents." Chap. 4 in *Frontiers of Business Cycle Research*, edited by Thomas F. Cooley. Princeton University Press.

Rotemberg, Julio J. 1982. "Sticky Prices in the United States." *Journal of Political Economy* 90(6): 1187–1211.

Sargent, Thomas J. 2015. "Robert E. Lucas Jr.'s Collected Papers on Monetary Theory." *Journal of Economic Literature* 53(1): 43–64.

Schechtman, Jack. 1976. "An Income Fluctuation Problem." *Journal of Economic Theory* 12(2): 218–41.

Shapiro, Matthew D., and Joel Slemrod. 2003. "Consumer Response to Tax Rebates." *American Economic Review* 93(1): 381–96.

Vissing-Jørgensen, Annette. 2002. "Limited Asset Market Participation and the Elasticity of

Intertemporal Substitution." *Journal of Political Economy* 110(4): 825–53.

Werning, Iván. 2015. "Incomplete Markets and Aggregate Demand." NBER Working Paper 21448.

Woodford, Michael. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press.

Yellen, Janet. 2016. "Macroeconomic Research after the Crisis." Speech at the Federal Reserve Bank of Boston, Boston, Massachusetts, October 14, https://www.federalreserve.gov/newsevents/speech/yellen20161014a.htm.

Yogo, Motohiro. 2004. "Estimating the Elasticity of Intertemporal Substitution when Instruments Are Weak." *Review of Economics and Statistics* 86(3): 797–810.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 195–214

Compensation and Incentives in the Workplace

Edward P. Lazear

Labor is supplied because most of us must work to live. Indeed, it is called "work" in part because without compensation, the overwhelming majority of workers would not otherwise perform the tasks. Of course, work can be rewarding and empowering, which adds to work's compensation. However, absent compensation, work would be much rarer and confined to those activities that are enjoyable, but not necessarily most needed by society.

The evidence that compensation affects worker behavior is overwhelming. At the most basic level, almost all of the labor component of GDP that is derived from work is paid rather than voluntary. Beyond that, the literature is full of examples where manipulating the pay structure alters worker behavior, by affecting either hours of work or output associated with it. Incentives are a necessary part of inducing the work that makes an economy go, even when those incentives must be self-imposed. Economists have understood the importance of incentives for decades; for example, they discussed it in the context of Soviet-style work environments.[1] The theory developed further during the 1970s and 1980s with modern agency theory, with early examples being Ross (1973), Lazear (1979, 1986), and Hölmstrom (1979).

[1] See Bergson (1944, 1978) and Weitzman (1984). Bergson discusses the importance of incentives within command economies towards the close of *The Structure of Soviet Wages: A Study in Socialist Economics* (1944, p. 204):

■ *Edward P. Lazear is the Davies Family Professor of Economics at the Graduate School of Business, and the Morris A. and Nona Jean Cox Senior Fellow at the Hoover Institution, Stanford University, Stanford, California.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.195 doi=10.1257/jep.32.3.195

Providing incentives can be important, indeed game-changing. In the study of Safelite Auto Glass installers discussed later (Lazear 2000b), a switch from hourly wage pay to a piece rate structure had an almost immediate and enormous effect of increasing productivity by 44 percent. The study showed both that changing a compensation scheme could have large effects and that economic theory does well in predicting these outcomes.

Personnel economics in general and the theory of incentives in particular has made its way into business. The combination of the academic literature, the popular press, and cohorts of students who have been schooled in the new approaches and who now are managers has influenced the way in which business is practiced. Some Silicon Valley companies like Success Factors and Merced Systems use these methods explicitly in providing expertise to other companies. Others, like Safelite mentioned above, and many other companies that use bonuses and promotions as motivators, incorporate the findings and analyses of incentive theory into their compensation practices. This is true not only in the United States, but also in Europe, particularly in Switzerland and Germany.

Forms of Incentive Pay

It is common to associate incentive pay with payment that is directly related to output.[2] This is too narrow. Virtually all pay methods provide incentives. A better taxonomy is to think of incentive compensation as being described by a two-by-three matrix, where columns relate to pay on input versus pay on output, and rows differentiate payment schemes as absolute payment that is discrete, absolute payment that is continuous, and payment that is primarily relative, as discussed in Lazear (2000c). Table 1 spells out the taxonomy.

For example, many workers face input-based pay with discrete incentives. The relevant input is that workers are paid per unit of time, either hours, weeks, months, or even years. Workers in this setting have no flexibility over the amount of time worked. For example, a retail clerk's contract may specify that 40 hours of work per week are required. The worker is paid per hour, but inflexibility on the choice of hours is part of the job. In contrast, a number of part-time jobs use input-based pay, but incentives are continuous in that the worker has choice over the amount

> "It is of great significance that Soviet equalitarianism was not of the utopian variety. That the worker requires a pecuniary incentive to acquire skill, to accept responsibility, to perform more arduous labor, and to increase his productivity, was an accepted principle of Soviet wage policy even in the period of War Communism. The equalitarian's appreciation of the magnitude of the incentive required may have been dim, particularly before 1920. But at least a sanguine appraisal of the conditions of supply set a lower limit to the reduction in differentials. If we may recur to the subject which was broached at the outset of this study, Soviet equalitarianism represented not an abandonment of capitalist wage principles, but at most a distorted application of them."

[2] These schemes are sometimes referred to as "high powered," as in Williamson (1975, 1985). This terminology is somewhat misleading because input-based, discrete schemes can also provide strong motivation to work, but load all incentives on achieving an exact target amount.

Table 1
Taxonomy of Incentive Compensation

	Payment on Input	*Payment on Output*
Discrete	Pay per hour with a specified hours requirement	Fixed payment for completion of construction project
Continuous	Time-based pay that allows worker choice of labor units supplied	Piece rates
Relative	Promotion tournaments based on subjective relative effort evaluation	Promotion tournaments based on some metric of relative output

of input supplied. For example, in academia, part-time faculty may be paid on the basis of the number of courses taught, which is an input measure (the output is what students derive from the course), but the instructors may be given some choice over how many courses they would like to teach. Although input-based contracts induce workers to put forth a particular amount of measured input, like hours, they may fall short in providing incentives for effort. As a result, time-based contracts are almost always coupled with some implicit or explicit performance standard that must be met to avoid being dismissed.

Common examples of continuous output-based schemes include piece-rate workers like those in agriculture who are paid according to the amount of crop harvested, salespersons whose compensation depends directly on sales, taxi drivers who rent their taxis for a flat rate and keep all revenue generated, and Uber and Lyft drivers. One major advantage of this approach is that it accommodates a variety of worker preferences. The scheme motivates those who want to work at high levels of effort as well as those who choose to work at lower levels of effort. A disadvantage is that a pure piece-rate scheme makes the worker bear risk associated with variations in exogenous factors like business conditions. Workers, especially low-wage ones, are less well-suited to bearing risk than are capital owners who can diversify their holdings.

Discrete output-based incentive schemes induce all workers to focus on a particular level of output. For example, a homeowner may hire a contractor to resurface a driveway at a given price. Payment for completion of the job motivates the contractor to perform, but all the incentives are concentrated on meeting the exact target—no more and no less. These all-or-nothing output-based contracts are less able to deal with heterogeneity, but they do create very strong incentives to get exactly the specified job done. Examples from the gig economy include Task Rabbit, where workers are paid a fixed amount to complete a specific task, or Upwork, which matches programmers with firms that need a specific piece of code to be written. Indeed, the rise of the gig economy may encourage a move away from time-based pay toward continuous versions of output-based pay, both because output is more easily measured in the gig economy and because hours worked are more difficult to measure.

Finally, both input- and output-based incentive schemes can be based on relative, rather than absolute performance. The classic form of relative scheme is a tournament in which the worker who does best receives a promotion. These schemes, discussed in more detail below, also provide incentives, even though they are both discrete and relative.

The theme of this essay is that incentives affect behavior and that economics as a science has made good progress in specifying how compensation and its form influences worker effort. This is a broad topic, and the purpose here is not a comprehensive literature review on each of many topics. Instead, a sample of some of the most applicable papers are discussed with the goal of demonstrating that compensation, incentives, and productivity are inseparably linked.

An underlying message of the discussion is that well-chosen compensation methods can affect positively both productivity and worker well-being. When firms provide an appropriate compensation structure, workers who join those firms benefit from being compensated at higher levels. Most of that additional compensation is inframarginal, meaning that the additional compensation more than offsets the disutility from the additional effort provided.

Piece Rates and Continuous Output Incentives

When piece rates are paid, some measure of output is specified and workers are paid on the basis of the number of units they produce. Piece-rate pay is best suited for situations in which output is easily observed and quality is not much of an issue (and can be ensured by occasional inspection). A standard example is agricultural harvesting, although even in agriculture, piece rates may be used more when crops are not delicate than where quality is more of an issue, as pointed out by Moretti and Perloff (2002). The literature is virtually unequivocal in documenting that for the circumstance where piece-rate pay is well suited, it provides incentives for workers to produce as predicted by standard theory.

In Lazear (2000b), mentioned above, I focus on Safelite Auto Glass installers, a company that switched from hourly wage pay to a piece-rate structure. Workers earned on average $11 per hour before the change to the piece rate and installed an average of 2.7 windshields per eight-hour day. The piece rate was set at approximately $20 per windshield, coupled with a minimum earnings guarantee. Productivity increased almost immediately by 44 percent. About half the increase is traceable to the workers who were present at the time of the switch, while the rest is attributable to the higher productivity of workers hired after the piece rate was put into place.

In some other prominent studies of piece rates, Shearer (2004) studies tree planters in British Columbia. He conducts a field experiment where nine randomly selected workers are observed for 120 days. During half the period, the workers were paid a fixed wage. During the other half, they were paid a piece rate. Again, both mean and variance are higher under piece rates than under fixed wages. Average output is about 21 percent higher under piece rates than under a fixed wage system. Bandiera, Barankay, and Rasul (2007) perform a field experiment where front-line supervisors, namely field managers on a fruit farm in the United Kingdom, were given a performance bonus halfway through the season. The field managers could increase fruit-picking productivity by their subordinates by working harder themselves, which involved clearing the filled crates faster; by assigning workers more strategically to rows of fruit; and by hiring better workers. The introduction of the performance bonus to field managers increased the overall productivity of their subordinates by about 25 percent.[3]

These examples of piece rates, although telling, of course do not mean that piece rates always result in higher output, compensation, and profit. In settings where output and/or quality are not very observable, piece rates may not work well. Even in settings well suited to piece rates, a piece rate that is set too high could raise cost per unit of output, or even reduce the level of output, if the income effect of the wage increase were large enough. But as a practical matter, firms are unlikely to put in place and maintain a piece-rate scheme that reduces profits.

Because worker output may be multidimensional and difficult to measure, a traditional piece-rate system is rarely used. This is the subject of some early work by Fernie and Metcalf (1998), which studies four firms in the United Kingdom, three of which are call centers and one of which is a bookmaker (a licensed betting office). The main thrust of this work is to compare the predictions of personnel economics to those of the older institutional literature with respect to pay schemes chosen. It is the choice of compensation scheme, rather than the effect of that choice on productivity, that is the subject of the analysis.

Performance-based pay, like piece-rate pay, seems to be associated with higher levels of output and pay for the average workers, but also with a higher dispersion of pay. As one example, Booth and Frank (1999) analyze performance-related pay using data from the British Household Panel Survey. Although they do not have information on output, they have detailed data on earnings and find that performance-related pay is associated with about 9 percent higher earnings for men and 6 percent higher earnings for women. In another study, Jirjahn and Kraft (2007) analyze the Hanover Panel of German Manufacturing Firms (1997). They have a variety of measures of compensation, including wage dispersion and the type of pay (piece rate or fixed wage) offered. They find that higher productivity at a firm is associated with higher wage dispersion. The elasticity is about .2, meaning that a 10 percent increase in wage dispersion is associated with 1.8 percent higher productivity. Additionally, they find that the productivity-dispersion elasticity is considerably higher in firms that operate an explicit piece-rate system, varying between .7 and .9, depending on the type of system used. The evidence on performance pay and piece rates suggests that when pay is overly compressed, it will tend to reduce incentives and output.

[3] In a study not focused on alternative methods of compensation (Lazear, Shaw, and Stanton 2015), we also find evidence of the importance of front-line supervisors on performance in the context of a company-based dataset on a technology-based services job, where the output of workers at a firm can be monitored by computer.

Team-Based Incentives

Consider a high-skill worker who also has the ability to help others develop their own skills. If that worker is paid a straight piece rate based on individual output, the worker would have no incentive to help others. However, if that worker is assigned to a relatively small team and paid partly on the basis of team output, the worker would face a tradeoff between spending time on personal work and spending time helping others. If compensation is based solely on the output of the team, the worker might even find it useful to spend most or all of the time helping others.

Joint production poses problems for incentive theory because workers may not have the right incentives to motivate their coworkers. Peer effects can interact with incentives and affect productivity, whether through monitoring, coaching, or motivation. In Kandel and Lazear (1992), my coauthor and I lay out the theory of the specific ways in which peer effects may operate, and Hölmstrom (1982) looks at free-riding in teams. An accumulating body of empirical evidence has validated a number of theoretical predictions from these models; for example, Mas and Moretti (2009) document positive, albeit small, peer spillover effects. (Substituting an above-average peer for a below-average peer increases a given worker's productivity by about 1 percent.)

Although team compensation suffers from free-rider effects, in small number settings, like medical practices where there are only a few physicians, incentive dilution may not be pronounced. Gaynor, Rebitzer, and Taylor (2004) find that incentives to conserve on costs, which increase physician take, are more effective in smaller physician groups—that is, groups in which the actions of individual physicians will have a greater effect on total savings.

But team incentives can prove useful in larger settings as well, perhaps by encouraging the adoption of more efficient methods of production. Employee stock option or ownership plans can be viewed as a broad-based team incentive. An obvious problem is that if a worker owns just a tiny fraction of the company, the return to the effort of that individual worker may be small relative to the cost. However, the ability to affect other workers magnifies the effect of effort on pay.

The effects of team incentives, although small, can be observed in real world data. Kruse (2016) reports that a review of about 100 studies suggests that there are positive effects on productivity, pay, job stability, and firm survival of employee ownership, although causation and interpretation questions remain. These results are consistent with those presented by Hamilton, Nickerson, and Owan (2003), who analyze a garment production plant that moved to team compensation in two stages, the first being voluntary and the second being compulsory. The authors find that average output rose by almost 20 percent after the switch to team production from individual piece-rate pay. In the context of this production process, if effort is defined more broadly to include effort in helping other workers on the job, a standard piece rate will fail to accomplish that effectively.

Bartel, Cardiff-Hicks, and Shaw (2017) study an international firm of lawyers in which compensation switched from pay on the basis of individual billings to one based on team revenues. Under the first system, the senior lawyers were less willing to share the work with their subordinates because their compensation depended strictly on what they billed themselves. After the switch, the same senior attorneys allocated more of the work to their subordinates.

Bandiera, Barankay, and Rasul (2005) explore the interaction between incentives and social networks. As in their other work, the data here come from studying agricultural workers on a UK farm. They compare the behavior of workers under piece rates, where compensation is independent of the performance of others, to that under a relative compensation scheme, where high output of a given worker imposes negative effects on the compensation of others because it raises the standard against which others are compared. Output is lower under the relative pay scheme. However, altruism does not seem the correct explanation. The disincentive of potentially high producers to cause negative externalities for others are internalized only when their actions can be monitored by the affected parties—specifically, when those who lose as a result of their efforts work alongside them and can observe the effort taken.[4] Their evidence demonstrates that compensation has significant effects on the way that workers behave toward one another.

Finally, some have documented how changing the pay structure relative to expectations might affect performance. Mas (2006) finds that performance changes when pay is exogenously altered from some reference point. Krueger and Mas (2004) find that at a Firestone tire factory, output quality declined (as evidenced by increased complaints about defective tires) following labor strife.

Relative Pay and Tournament Theory

"Tournament theory" is the name used to describe the literature that focuses on providing incentives to workers on the basis of their relative performance within a firm. The title is not much of an abstraction from the way many workers—especially those climbing the corporate ladder—think about their situations.

To grasp the intuitition behind this approach, consider two players competing in a tennis match—say, the Wimbledon finals. The two players are (presumably) fairly close in skill in an absolute sense, but even when the match is very close, the prize for the winner is considerably larger than that for the loser. The winner's prize is based on relative, not absolute, performance. There is an optimal spread between the prize for the winner and the runner-up, and optimality is defined so as to elicit the efficient level of effort, not the maximum effort possible. For example, it might be possible to set the prize so high that players would be willing to risk death or try to kill the other player to win, as was the case in gladiatorial tournaments, but that would not be efficient because the added output would not cover the cost.

Within a firm, receiving a promotion is akin to winning a tennis tournament. The contestants in this case are typically managers at a lower level in the firm's

[4] An older literature documents how workers may react to "rate-busters," defined as workers who perform at high levels in an industrial setting and thus have the effect of lowering compensation per unit of output for all others. For an example, see Roy's (1952, 1954) studies of a machine shop.

hierarchy. Promotion decisions almost always require relative rankings. The raise associated with a promotion motivates young associates at consulting firms to exert high levels of effort in order to win a promotion to partner, just as the hope of tenure motivates assistant professors in academia. However, if the value of winning a promotion is too high, the output from the additional incentives provided will fall short of the additional wages that must be paid to induce workers to accept the job at the outset. The difference between the spread in wages for the promoted and the unpromoted should be just large enough to induce the efficient level of effort, but not so high as to exceed it.

These intuitive ideas have been derived formally in early game-theoretic analyses (Lazear and Rosen 1981; Green and Stokey 1983; Nalebuff and Stiglitz 1983), and demonstrated in laboratory experiments (for instance, Bull, Schotter, and Weigelt 1987; Falk, Fehr and Huffman 2008) and in sports environments (for an example from golf, see Ehrenberg and Bognanno 1990).[5]

In the more germane context of business, Eriksson (1999) finds that many implications of tournament theory hold when looking at data on about 2,600 executives in 210 Danish firms. First, the larger the number of candidates for a promotion, and thus the smaller the probability of being promoted, the higher the pay jump associated with that promotion, which is necessary to provide incentives. Second, the jump in pay for promotions at high levels is greater than that at low levels. This is implied by the theory because part of the reward for low-level promotions is the option value of obtaining higher-level additional promotions (Rosen 1976).

The second finding is also corroborated by Belzil and Bognanno (2008), who examine 600 US firms having 25,000 executives. Like Eriksson, they find that pay jumps are larger for promotions near the top than near the bottom. Classifying levels within firms into nine categories, a promotion from the bottom to the next level generates a 15 percent raise, whereas a promotion from the second-to-the-top level to the top generates a 94 percent raise.

Effort is difficult to observe in data, but absenteeism can, in some situations, be used as a proxy for the effort that workers are willing to commit to a job. Drago and Garvey (1998) examine data from 23 Australian firms. They find that the larger the spread in wages between workers and supervisors, the less absenteeism. They interpret this as workers being willing to commit higher effort to the job when the incentives from promotion to a higher rank are larger.

There are indirect tests of incentives to put forth effort that use observable outcomes as a proxy for effort. The Danish data used by Eriksson (1999) provide a test of outcomes in tournament settings. Eriksson finds that those firms that award larger raises to promotion have better outcomes as measured by profits—and also pay higher average wages. Using US data, Kale, Reis, and Venkateswaran (2009)

[5] Gneezy, Niederle, and Rustichini (2003) explore the way men and women respond to piece-rate and tournament structures. They find no statistically significant difference by gender in response to piece rates, but men outperform women in tournament incentive settings. Noteworthy is that women's performance is higher in single-sex tournaments than in a piece-rate treatment.

find that the difference between the pay of chief executive officers and vice presidents is strongly and positively related to return on assets, return on equity, and market-to-book value. Using data from the Swedish Registry, which surveys about 10,000 managers, Heyman (2005) finds a positive and significant effect of intra-firm wage dispersion on profits and average pay. He also finds that dispersion tends to be larger in firms that experience higher market demand volatility, which is consistent with tournament predictions. When workers compete in a noisy environment, larger incentives and wage dispersion are required to compensate for the diluting effects on incentives of higher risk. These studies document that performance varies in a way consistent with tournament theory, and they provide indirect evidence that effort is affected by altering compensation structures in the ways predicted by tournament theory.

Various other findings also support the tournament view of incentives. Mobbs and Raheja (2011) find that firms that have multiple candidates competing for promotion among top executives do better than those with a groomed successor. Actually, the prediction is that the relationship between number of candidates and effort exerted should be an inverted U. If there is a groomed successor, the potential competitors are not highly motivated to compete for the job, just as the authors find. But with an extremely large number of potential competitors, incentives to put forth effort would be reduced as well, because the change in probability of winning the tournament/promotion from exerting additional effort becomes very small.

Compensation may affect managers' risk-taking behavior. The classic concern is that risk-averse managers adopt overly safe strategies to protect their jobs. Because shareholders can diversify their portfolios, managers who implement the desires of the principals should behave as if they are risk neutral. Tournaments may encourage more risk taking among risk-averse managers because winning a multicontestant tournament is a tail-event. A sports parallel is that, in Olympic downhill skiing, a cautious approach will assure a loss of the gold medal, even if it minimizes expected time to complete the course. Winning requires great talent, but also a high draw of positive luck. Choosing low-variance strategies precludes getting a large positive draw on luck.

Properly structured stock options can also be used to address managerial risk aversion. Just as was the case in the downhill skiing analogy, call options create incentives for adopting riskier strategies (as noted in Jackson and Lazear 1991). If the exercise price of the option is high, then only a manager who adopts a high-risk strategy will end up with a stock that is "in the money," that is, has a value higher than the exercise price.[6] Kini and Williams (2012) find that increasing tournament-like

[6] An intriguing question sometimes posed in the compensation literature is whether firms might wish to grant put options, rather than traditional call options. With a traditional stock call option, managers have incentive to put forth effort that translates into higher stock prices, because then they can sell the stock at a higher price differential over the exercise price of the option, which is given and predetermined. But consider an alternative. Managers could be given higher base salaries and required to short put options on stock. If the stock price falls below a certain level, managers would have to purchase company stock at a price higher than the market price, forcing them to overpay for shares of their company's own stock. Both kinds of options provide incentives. The difference is that even risk-neutral managers who are short

incentives increases risk taking. To the extent that the losing prize—the wage of the unpromoted worker—is relatively constant, an increase in spread is akin to increasing variance of an option. Kini and Williams find that a bigger pay gap at a firm implies more cash flow volatility and more volatility in returns. Additionally, firms with bigger pay gaps undertake riskier investments (like emphasizing research and development over tangible assets) and have higher leverage ratios.

Not all incentives associated with more risk taking are positive. Hass, Muller, and Vergauwe (2015) find that a larger variance in the pay structure creates more dysfunctional responses like fraud, as reflected in class-action lawsuits against the firm. More generally, when a worker's promotion depends on relative performance, workers have an incentive to be uncooperative with fellow workers and in the extreme, even to sabotage their efforts. Drago and Garvey (1998) find that workers are less willing to be helpful to co-workers when the promotion rewards are greater. Falk, Fehr, and Huffman (2008) document the same behavior in a lab setting, and even document a rise in sabotage that increases with the prize spread. Additionally, they find that the pay setter compresses pay in response to this. Thus, a tradeoff arises here, as predicted in Lazear (1989). Reducing the spread between the promoted and the unpromoted has adverse consequence for good effort, but mitigates the uncooperative behavior that is inherent in any relative compensation scheme. The dangers of sabotage or uncooperative behavior are another reason for limiting the size of the promotion incentives. In real-world settings, compensation-induced incentives can end up being beneficial or adverse, when poorly designed. But either answer confirms an underlying theme of this essay, which is that compensation schemes affect worker incentives and behavior, both in theory and in practice.

Career Incentives and the Experience–Earnings Profile

For many workers, promotion is not a realistic possibility. For example, consider a middle-aged manager who has been in the same position for many years with little hope of moving up in the hierarchy. Other than paying the worker directly on output, which may not be easily observed, it may be difficult to motivate workers in these positions. For such workers, the experience–earnings relationship can serve as a way to motivate workers who are neither paid piece rates nor are candidates for promotion.

The theory behind these schemes was exposited in Lazear (1979, 1981). The idea is rooted in the empirical observation that experience–earnings profiles tend to be nonnegatively sloped throughout a worker's career. This pattern holds despite both casual and more serious empirical evidence suggesting that productivity declines with age. For example, age at some point brings a decline in the ability to conduct logical thinking and reasoning (Ruth and Birren 1985), a diminution of

a put option would want to choose safe strategies to avoid a big downside effect on stock price whereas managers who are long a call option would want to take more risk. The fact that firms make managers long on call options and rarely (if ever) short on put options suggests that shareholders and the boards are trying to increase managerial risk taking, rather than decrease it.

creativity (Florida 2002), and declining abilities to store and process information, solve problems, deal with complexity, and adjust to new situations (Kaufman and Horn 1996; Ryan, Sattler, and Lopez 2000).

Young and middle-aged workers often produce more than the more senior workers who receive higher compensation. How can this pattern survive in equilibrium? The theory argues that efficiency is enhanced by underpaying the young and overpaying the old. Middle-age and senior workers have incentives to keep their levels of effort high because they earn rents on the job relative to what they would receive if they were to be forced to leave the job. The scheme requires some monitoring, but this monitoring can occur occasionally and stochastically.

The theory has a number of testable implications. When first laid out, the goal was to provide an economic rationale for mandatory retirement. Because senior workers are overpaid relative to their productivity, they want to work longer than is efficient, which can be defined as retiring when the value of their leisure exceeds their productivity at work.

Because it is difficult in most settings to observe the experience–productivity relationship, direct tests on the slope of experience–earnings versus experience–productivity profiles are virtually nonexistent.[7] In an early paper, Hutchens (1987) uses a clever indirect method that relies on the National Longitudinal Survey combined with the Dictionary of Occupational Titles in order to "test whether jobs that involve repetitive tasks tend to be characterized by an absence of pensions, mandatory retirement, long job tenures, and high wages for older workers." The hypothesis is that jobs with technologies that are well-suited to supervision or oversight can use more direct monitoring schemes and at the extreme, output-based pay, to motivate workers. For those where the job tasks are less easily measured, one would expect to observe more pensions, long tenures, mandatory retirement, and wages that increase more rapidly. He finds that workers who perform nonrepetitive tasks are 9 percent more likely to have a pension, have wages that are 28 percent higher (which, because of controls, reflects a steepening of the experience–earnings profile), and an 18 percent longer job tenure.

In another test, Goldin (1986) argues that women historically had shorter attachments to the labor force and as a result, should be more likely to have direct incentive pay and less likely to have experienced-based incentive pay. She documents that women were more likely to be on piece rates than men, although over time, the difference between men and women should disappear as female labor force behavior resembled that of men. Indeed, other authors have found that to be the case. Booth and Frank (1999) using British data find that women were actually less likely to be on piece rates.

How might laws against age discrimination affect this experience–earnings relationship? Neumark and Stock (1999) use variation in state and federal age discrimination laws to show that stiffer laws against age discrimination actually lead

[7] The Safelite data allow for an explicit test of the theory because productivity is measured precisely. The learning curve was quite steep for the installers, but wages continued to grow beyond the point at which experience effects on productivity flattened (Lazear 1999).

to steeper experience–earnings profiles. Their interpretation is that age discrimination laws make it harder to fire older workers, which means that a firm will find it more difficult to renege on its promise to pay older workers more. In this sense, age discrimination laws make the firm's promise to pay experienced workers high salaries credible for younger cohorts, and experience-based incentives are more likely to be provided after age discrimination laws are passed.

More recent evidence is provided by Huck, Seltzer, and Wallace (2011). They use an experimental setting and find that when firms cannot commit successfully, there is "breakdown of worker-firm relations and a dramatic loss in efficiency … It is this comparison that really underlines the success of the Lazear model—the difference deferred compensation makes."

Labor Supply and Incentives for Inputs

Incentive theory is less frequently applied to or discussed in the context of payment for input, but it is just as relevant. At the most basic level, determining the elasticity of labor supply is fundamentally a question of incentives. This is not the place to attempt a survey of a vast literature on labor supply, but a few issues are directly relevant to the study of compensation incentives. First, raising wages induces individuals to work more. Although the elasticity of labor supply estimates vary from very small to substantial, depending on the group and margin considered, most of the literature supports the view that hours worked and labor force participation respond positively to pay. Second, government programs, which exogenously alter the benefit from work, affect the amount of work in an economy.

Among the earliest papers to study the impact of exogenous wage changes on worker behavior is Rosen (1976), who examines whether tax changes that affected take-home wages led to a change in labor supply. Rosen's main conclusion based on the Survey of Work Experience for Women 30–44 (1967) is that married women's labor supply is highly responsive to tax changes that affect the take-home wage rate. Another analyses of the effects of tax changes on labor supply is Eissa and Hoynes (2004), which studies the effect of the Earned Income Tax Credit on behavior, with a focus on labor force participation. One finding is, "A $1 increase in the net wage raises the likelihood that wives work by 2.7 percentage points, or 4.2%."

Yet another example of exogenous government programs that affect worker behavior is disability insurance. In an early study, Gruber (2000) compared work and disability behavior in different Canadian provincial regimes. Changes made in 1987 in the benefits offered by provinces other than Quebec allow Gruber to conduct a differences-in-differences analysis. Gruber reports that a 36 percent rise in disability benefits resulted in a rise in nonemployment of 11.5 percent from the baseline value. Recent analysis from 2006–2016 with US data reveals a correlation of .97 between the number of new applications for disability insurance (annual) and the unemployment rate (Lazear 2017). In this case, the interpretation is that when a recession occurs, incentives for work are reduced, and some who had qualifying disabilities will exercise their rights to receive disability insurance, forgoing work.

Other Views and Questions

A wide range of empirical evidence demonstrates that the form and level of workplace compensation has important effects on the incentives and behavior of workers and managers. Two additional questions are relevant. First, what is the role of nonmonetary incentives and of intrinsic motivation? Second, are there situations in which market forces may systematically provide incentives that are too large or too small for efficiency? The issue of overcompensation of chief executive officers and other high-level managers is a special case of providing too strong or perhaps the wrong kind of incentives.

Nonmonetary Incentives and Intrinsic Motivation

The fact that workers are motivated not only by money is neither novel nor controversial. The idea dates back to Adam Smith (1776, Bk. 1, Ch. X) and falls under the general rubric of "compensating differentials."[8] Rosen (1974) modernizes the idea and derives the market equilibrium, understanding that the firm's profit incentives interact with worker preferences to determine the price that workers must pay to have their tastes accommodated. Rosen's insights imply that attributes that could be provided at no cost to the firm would not carry with them a compensating differential, even if workers had preferences for the attributes.

Consider intrinsic motivation. Some work occurs in the complete absence of pay. Charitable contributions equal about 2 percent of GDP (Zinsmeister 2016; Giving USA 2017), but that is not the same as work done without pay. Most charitable contributions go to organizations like those in health care research that hire workers in a competitive labor market and pay wages. Volunteer work is a more direct measure of the amount of work done without monetary compensation and presumably based on intrinsic motivations, which might include altruism or group identification. However, its value is only in the range of 1 percent of GDP (as reported by Hrywna 2017).

Household work accounts for much more work that is unpaid, at about 25 percent of GDP (Bridgman, Dugan, Lal, Osborne, and Villones 2012). However, even though household work is unpaid, it is associated with an almost ideal incentive structure because it is akin to self-employment, where a high proportion of the rewards to effort are captured either by the provider of the services or by immediate family members. The provider of household services is the residual claimant of services consumed. Services consumed by other family members deliver first-best incentives when altruism by the provider is sufficient to make him or her view services consumed by family members as equal in value (on the margin) to those

[8] For example, see Book I, Chapter X of Adam Smith's (1776) *Wealth of Nations*: "The five following are the principal circumstances which, so far as I have been able to observe, make up for a small pecuniary gain in some employments, and counter-balance a great one in others: first, the agreeableness or disagreeableness of the employments themselves; secondly, the easiness and cheapness, or the difficulty and expence of learning them; thirdly, the constancy or inconstancy of employment in them; fourthly, the small or great trust which must be reposed in those who exercise them; and fifthly, the probability or improbability of success in them."

consumed directly. Relatedly, some consumption might be thought of as work. For example, personal gardening produces output that is not counted in GDP. It is possible that gardening as a consumption activity or cooking for enjoyment fails to be counted in cited estimates of household work. If that is the case, then the estimates (for example, BEA Blog 2012) understate the true share of GDP that is unpaid.

Is intrinsic motivation of broad importance for standard labor market settings? An influential paper by Deci (1972) argued that monetary incentives can in some cases crowd out intrinsic motivation, where people take actions because those actions are personally fulfilling. One famous example is that blood donations declined when the collecting agency moved from voluntary to paid contributions (Titmuss 1970). More germane to the subject at hand is the work of Bruno Frey and coauthors, who have critiqued monetary incentives as a mechanism for enhancing worker productivity, and pointed to some cases in which making explicit the reward has adverse effects on work effort (for example, Frey and Oberholzer-Gee 1997; Frey 1997).

Work that is motivated intrinsically, say by altruism as in the case of work to benefit family members, clearly exists. Nonetheless, most work is done to obtain the direct pay that results from it. At least on the margin, it is necessary to compensate individuals to get them to put forth effort. The reason is straightforward. If effort has value in the market or specifically to a particular employer, then the employer will be willing to pay for that additional effort. That will push the worker to the point where intrinsic motivation has run out and where it is necessary to pay workers to elicit additional effort, a point made as early as Marshall's (1890) landmark textbook.

For example, many people volunteer some hours, but there are only so many hours that are supplied without pay. Even socially oriented individuals must be paid to exceed that limit. Some salespeople might work a few hours just for the intrinsic joy of selling, but it is unimaginable that much sales work would be done absent explicit compensation. The studies of piece rates and how they motivate workers provide evidence to this effect.

Indeed, the role of compensation and management in general is to induce workers to take those actions that are of value to the firm but would not be provided voluntarily. Much of management is about inducing workers to perform those tasks that have value but are not intrinsically rewarding. Volunteer work and effort provided through intrinsic motivation is consistent with standard economic theory. Economists understand equilibrium and on the margin, a dollar spent raising productivity one way must be as effective as a dollar spent raising it any other way. A firm that has exhausted its profitable means of raising productivity through changes in the compensation structure might find that it has opportunities to raise productivity even further by appealing to or changing worker tastes. The converse holds as well. A firm that has the ability to create a positive culture in the firm must balance the cost of improving culture with simply paying workers to put forth additional effort. An optimizing firm allocates expenditures such that the marginal dollar spent on each form of compensation produces the same worker satisfaction.

Although economists have little to add to psychological theories of preferences, economic theory does provide a guide as to when this kind of psychological manipulation will more plausibly have effects on incentives and pay—and when not.

Altering worker preferences to further align their incentives with those of the firm is likely to be concentrated on those higher-level and higher-paid workers who have stronger attachment to the firm. Job churn in the United States totals about 60 million jobs per year, which is about two-fifths of the entire worker force, according to the Bureau of Labor Statistics (2017) Job Openings and Labor Turnover Survey (JOLTS). Many workers have tenuous connection to their firm, so the firm will have little incentive to invest in ways that would affect their preferences.

Additionally, one should always consider the salary level in question. A manager who earns $300,000 per year may prefer a $5,000 improvement in working conditions over $5,000 in pay. A worker who earns $25,000 per year is less likely to prefer $5,000 in more appealing working conditions to $5,000 in pay. Working conditions are a normal and likely a luxury good, which implies that extrapolation of responses from one group to another is dangerous. This would also cut across occupations. Sorting into occupations is not random, and some occupations are chosen by persons who have an inclination toward nonmonetary rewards. Obvious cases involve missionary work, which implies significant hardship and very low compensation. Those who choose those occupations value social rewards over monetary ones to a greater extent than, say, investment bankers. Less extreme cases are abundant in the labor market.

Beyond these issues, firm specificity matters. In order for a firm to be willing to take an action that affects a worker's productivity, in this case bearing costs to change preferences, the firm must be able to capture the returns from that investment, which is only possible if the value produced is firm specific. Were the worker not partial to the current firm, any investment that made the worker more valuable, either because skills were enhanced or because the worker was willing to supply more effort at a given wage, could not survive competition. The worker might be willing to supply more effort at the same price, but if that preference had value to other firms in the marketplace, they would bid up the wage commensurate with the additional effort and the firm that made the initial investment would not recoup any of it. As a consequence, firms can only be expected to undertake investments that change worker preferences when those workers are already earning rents at the current firm relative to others or when the change in preferences from the investment creates a stronger attachment to the current firm directly.

Many of the critiques of compensation incentive theory highlight anomalies, frequently arising from results in laboratory studies. My personal view of this literature, discussed in Lazear (2000a), is that criticism should be welcome. It is dangerous to be excessively smug about our own theories and ideas, and those who force us to reexamine our analyses and predictions provide a useful service. But that does not imply that the discipline and framework provided by standard economic theory should be rejected. The goal of science is to make sense out of nonsense, not the reverse, and it is incumbent on scholars to think deeply enough so as to incorporate into or reformulate theory to provide a unified picture. Scattered facts and inconsistencies do not form a scientific literature, but they may help stimulate science to be better.

Sometimes, results from laboratory experiments mislead because they lack a well-posed question. Consider studies that purport to compare piece rates with a tournament pay scheme. To determine which is more effective (holding the environment constant), it is necessary to compare an appropriately designed piece rate against an appropriately designed tournament. It is always possible to make a piece rate more effective than a tournament by increasing the piece rate and reducing the spread between the winning and losing prizes in a tournament. The reverse is also true. Changes in environment can affect the outcome systematically, too. If experiments are to offer useful lessons, they must take considerable care to follow the scientific method, as used by economists and physical scientists alike. Lavy (2002) offers a fine example of how this can be done. He examines a given amount of money spent on teachers and compares that to a similar amount spent on more general educational resources. Lavy looks at whether a dollar spent on teacher pay is more effective on student outcomes than a dollar spent on books and other resources that a school has, but which do not go to teachers directly. Holding costs constant in making comparisons between compensation schemes, while seemingly obvious, is an essential ingredient of experimental design.

Markets and the Amount of Compensation: The Case of Managerial Pay

The literature on pay structures suggests that relative compensation often provides positive motivation. But many are shocked by the high levels of managerial compensation, particularly for CEOs, prevalent in the United States. Are these high levels of pay justified? Do they provide appropriate incentives and attract needed talent?

There is an economic literature that criticizes the pay of CEOs and other high-level executives, claiming that it is too high. Overcompensation is attributed to the CEOs ability to capture the board of directors or to social pressure to pay salaries commensurate with some norm. It is argued that compensation at the top of the firm is pushed beyond its efficient level, creating too large a spread in the tournament prize structure. Stated analytically, the argument is that even if the high compensation results in additional effort or better talent, the compensation premium does not produce enough additional value to cover its cost.

Research on the relation of managerial compensation to social factors attracted attention in the late 1980s. O'Reilly, Main, and Crystal (1988) presented evidence that the pay of chief executive offers was correlated with the compensation of directors. Their favored interpretation was that the norms for compensation come at least in part from introspection, and directors use their own pay as a frame of reference when setting compensation for a top corporate officer. The implication is that some chief executive officers may be overpaid simply because they are fortunate in having a board with highly paid directors. Another interpretation, however, is that this correlation is created by sorting. Top companies recruit top directors who are highly paid, both in their regular jobs and as board members. The same companies also recruit the most talented chief executive officers, and the market for such talent gives them high pay as well. If this is true, the pay of chief executive officers and director at a firm would be correlated, but that correlation could be appropriate and profit-enhancing.

A large literature connects pay of chief executive officers to firm size, and most of that literature argues that the connection can be justified. The most talented manager is more effectively utilized when running a multi-billion dollar company than when running the local hardware store. If a very able chief executive officer increases the value of a company by 10 percent more than that of an average alternative, this will have a larger effect in companies that have more capital (Rosen 1976; Gabaix and Landier 2008, and the references therein).

In yet another critique, Bebchuk, Cremer, and Peyer (2011) find evidence that the larger the share of chief executive officer pay among the top five executives in the firm, the lower the profit and efficiency that result. Their preferred interpretation is that the excessive share going to the chief executive officers does not enhance shareholder nor worker value. Perhaps, but this pattern could also mean that other executives are underpaid. Alternatively, the pattern could arise from a form of survivorship bias in the data. Economic theory suggests that those firms that give their chief executive officers *either* too high or too low a share will have lower profits than those who pay the optimum share. Undercompensated chief executive officers are more likely to be lured away by the competition. Overcompensated chief executive officers lower the profitability of the firm, but that factor alone is unlikely to drive a firm out of business because compensation for chief executive officers is generally a small part of total costs to the firm. But because of this survivorship bias, there will be fewer data points below the optimum than above, even if there is no built-in bias toward overcompensation.

Setting pay for corporate executives is an inexact science applied in a continually evolving corporate landscape, so it is unsurprising that some chief executive officers will appear overpaid or underpaid, especially when the case is viewed in hindsight. But the evidence that chief executive officers receive pay that exceeds that which is consistent with profit maximization is neither unequivocal nor compelling. At the same time, the literature that presents evidence that rationalizes the high pay of chief executive officer and other managers, while of high quality, remains limited.

Conclusion

Compensation and its structure have profound effects on worker motivation. The theory of incentives laid out decades ago and refined in recent years continues to garner support as more firm-based data have become available.

■ *I am grateful to Erielle Davidson for her extensive work reviewing the literature that is discussed in this essay.*

References

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *Quarterly Journal of Economics* 120(3): 917–62.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2007. "Incentives for Managers and Inequality among Workers: Evidence from a Firm-Level Experiment." *Quarterly Journal of Economics* 122(2): 729–70.

Bartel, Ann P., Brianna Cardiff-Hicks, and Kathryn Shaw. 2017. "Incentives for Lawyers—Moving Away from 'Eat What You Kill.'" *Industrial and Labor Relations Review* 70(2): 336–58.

BEA Blog. 2012. "What is the Value of Household Work?" June 11. https://blog.bea.gov/2012/06/11/household-work/.

Bebchuk, Lucian A., K. J. Martijn Cremers, and Urs C. Peyer. 2011. "The CEO Pay Slice." *Journal of Financial Economics* 102(1): 199–221.

Belzil, Christian, and Michael Bognanno. 2008. "Promotions, Demotions, Halo Effects, and the Earnings Dynamics of American Executives." *Journal of Labor Economics* 26(2): 287–310.

Bergson, Abram. 1944. *The Structure of Soviet Wages: A Study in Socialist Economics.* Cambridge, MA: Harvard University Press.

Bergson, Abram. 1978. "Managerial Risks and Rewards in Public Enterprise." *Journal of Comparative Economics* 2(3): 211–25.

Booth, Alison L., and Jeff Frank. 1999. "Earnings, Productivity, and Performance-Related Pay." *Journal of Labor Economics* 17(3): 447–63.

Bridgman, Benjamin, Andrew Dugan, Mikhael Lal, Matthew Osborne, and Shaunda Villones. 2012. "Accounting for Household Production in the National Accounts, 1965–2010." *Survey of Current Business* 92(5): 23–36.

Bull, Clive, Andrew Schotter, and Keith Weigelt. 1987. "Tournaments and Piece Rates: An Experimental Study." *Journal of Political Economy* 95(1): 1–33.

Bureau of Labor Statistics. 2017. *Job Openings and Labor Turnover Survey.* Washington, D.C.: United States Department of Labor, Bureau of Labor Statistics. https://www.bls.gov/jlt/.

Deci, Edward L. 1972. "Intrinsic Motivation, Extrinsic Reinforcement, and Inequality." *Journal of Personality and Social Psychology* 22(1): 113–20.

Drago, Robert, and Gerald T. Garvey. 1998. "Incentives for Helping on the Job: Theory and Evidence." *Journal of Labor Economics* 16(1): 1–25.

Ehrenberg, Ronald G., and Michael L. Bognanno. 1990. "Do Tournaments Have Incentive Effects?" *Journal of Political Economy* 98(6): 1307–23.

Eissa, Nada, and Hilary Williamson Hoynes. 2004. "Taxes and the Labor Market Participation of Married Couples: The Earned Income Tax Credit." *Journal of Public Economics* 88(9–10): 1931–58.

Eriksson, Tor. 1999. "Executive Compensation and Tournament Theory: Empirical Tests on Danish Data." *Journal of Labor Economics* 17(2): 262–80.

Eriksson, Tor, Sabrina Teyssier, and Marie-Claire Villeval. 2009. "Self-Selection and the Efficiency of Tournaments." *Economic Inquiry* 47(3): 530–48.

Falk, Armin, Ernst Fehr, and David Huffman. 2008. "The Power and Limits of Tournament Incentives." March 28, http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=F93C0F6D4848B10E29CA25D037602B3F?doi=10.1.1.724.7243&rep=rep1&type=pdf.

Fernie, Sue, and David Metcalf. 1998. "(Not) Hanging on the Telephone: Payment Systems in the New Sweatshops." CEP Discussion Paper 390, Centre for Economic Performance, London School of Economics and Political Science.

Fernie, Sue, and David Metcalf. 1999. "It's Not What You Pay it's the Way that You Pay it and that's What Gets Results: Jockeys' Pay and Performance." *Labour* 13(2): 385–411.

Florida, Richard. 2002. *The Rise of the Creative Class: And How It's Transforming Work, Leisure, Community and Everyday Life.* New York: Basic Books.

Frey, Bruno S., and Felix Oberholzer-Gee. 1997. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out." *American Economic Review* 87(4): 746–55.

Frey, Bruno S. 1997. *Not Just for the Money: An Economic Theory of Personal Motivation.* Cheltenham, UK: Edward Elgar.

Gabaix, Xavier, and Augustin Landier. 2008. "Why has CEO Pay Increased So Much?" *Quarterly Journal of Economics* 123(1): 49–100.

Gaynor, Martin, James B. Rebitzer, and Lowell J. Taylor. 2004. "Physician Incentives in Health Maintenance Organizations." *Journal of Political Economy* 122(4): 915–31.

Giving USA. 2017. "Giving USA 2017: Total Charitable Donations Rise to New High of $390.05 Billion." Press Release. *Giving USA,* June 12. https://givingusa.org/giving-usa-2017-total-charitable-donations-rise-to-new-high-of-390-05-billion/.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in Competitive Environments: Gender Differences." *Quarterly Journal of Economics* 118(3): 1049–74.

Goldin, Claudia. 1986. "Monitoring Costs and Occupational Segregation by Sex: A Historical Analysis." *Journal of Labor Economics* 4(1): 1–27.

Green, Jerry R., and Nancy L. Stokey. 1983. "A Comparison of Tournaments and Contracts." *Journal of Political Economy* 91(3): 349–64.

Gruber, Jonathan. 2000. "Disability Insurance Benefits and Labor Supply." *Journal of Political Economy* 108(6): 1162–83.

Hamilton, Barton H., Jack A. Nickerson, and Hideo Owan. 2003. "Team Incentives and Worker Heterogenity: An Empirical Analysis of the Impact of Teams on Productivity and Participation." *Journal of Political Economy* 111(3): 465–97.

Hass, Lars Helge, Maximilian A. Müller, and Skrålan Vergauwe. 2015. "Tournament Incentives and Corporate Fraud." *Journal of Corporate Finance* 34: 251–67.

Heyman, Fredrik. 2005. "Pay Inequality and Firm Performance: Evidence from Matched Employer-Employee Data." *Applied Economics* 37(11): 1313–27.

Hölmstrom, Bengt. 1979. "Moral Hazard and Observability." *Bell Journal of Economics* 10(1): 74–91.

Hölmstrom, Bengt. 1982. "Moral Hazards in Teams." *Bell Journal of Economics* 13(2): 324–40.

Hrywna, Mark. 2017. "Volunteer Time Value Pegged at $193.0 Billion." *NonProfit Times,* April 20. http://www.thenonprofittimes.com/news-articles/volunteer-time-value-pegged-193-billion/.

Huck, Steffen, Andrew J. Seltzer, and Brian Wallace. 2011. "Deferred Compensation in Multi-period Labor Contracts: An Experimental Test of Lazear's Model." *American Economic Review* 101(2): 819–43.

Hutchens, Robert M. 1987. "A Test of Lazear's Theory of Delayed Payment Contracts." *Journal of Labor Economics* 5(4): S153–S170.

Jackson, Matthew O., and Edward P. Lazear. 1991. "Stock, Options, and Deferred Compensation" In *Research in Labor Economics,* vol. 12, edited by Ronald G. Ehrenberg, 41–62. Greenwich, CT: JAI Press.

Jirjahn, Uwe, and Kornelius Kraft. 2007. "Intra-Firm Wage Dispersion and Firm Performance: Is There a Uniform Relationship?" *Kyklos* 60(2): 231–53.

Kale, Jayant R., Ebru Reis, and Anand Venkateswaran. 2009. "Rank-Order Tournaments and Incentive Alignment: The Effect on Firm Performance." *Journal of Finance* 64(3): 1479–1512.

Kandel, Eugene, and Edward P. Lazear. 1992. "Peer Pressure and Partnerships." *Journal of Political Economy* 100(4): 801–817.

Kaufman, Alan S., and John L. Horn. 1996. "Age Changes on Tests of Fluid and Crystallized Ability for Women and Men on the Kaufman Adolescent and Adult Intelligence Test (KAIT) at Ages 17–94 Years." *Archives of Clinical Neuropsychology* 11(2): 97–121.

Kini, Omesh, and Ryan Williams. 2012. "Tournament Incentives, Firm Risk, and Corporate Policies." *Journal of Financial Economics* 103(2): 350–76.

Krueger, Alan, and Alexander Mas. 2004. "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires." *Journal of Political Economy* 112(2): 253–89.

Kruse, Douglas. 2016. "Does Employee Ownership Improve Performance?" IZA World of Labor. December. https://wol.iza.org/articles/does-employee-ownership-improve-performance/long.

Lavy, Victor. 2002. "Evaluating the Effects of Teachers' Group Performance Incentives on Pupil Achievement." *Journal of Political Economy* 110(6): 1286–1317.

Lazear, Edward P. 1979. "Why Is There Mandatory Retirement?" *Journal of Political Economy* 87(6): 1261–84.

Lazear, Edward P. 1981. "Agency, Earnings Profiles, Productivity, and Hours Restrictions." *American Economic Review* 71(4): 606–20.

Lazear, Edward P. 1986. "Salaries and Piece Rates." *Journal of Business* 59(3): 405–431.

Lazear, Edward P. 1989. "Pay Equality and Industrial Politics." *Journal of Political Economy* 97(3): 561–80.

Lazear, Edward P. 1999. "Personnel Economics: Past Lessons and Future Directions." Presidential Address to the Society of Labor Economists, delivered in San Francisco on May 1, 1998. *Journal of Labor Economics* 17(2): 199–236.

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

Lazear, Edward P. 2000b. "Performance Pay and Productivity." *American Economic Review* 90(5): 1346–61.

Lazear, Edward P. 2000c. "The Power of Incentives." *American Economic Review* 90(2): 410–14.

Lazear, Edward P. 2017. "The Incredible Shrinking Workforce." *Wall Street Journal,* December 7. https://www.wsj.com/articles/the-incredible-shrinking-workforce-1512692004.

Lazear, Edward P., and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy* 89(5): 841–64.

Lazear, Edward P., Kathryn Shaw, and Chris Stanton. 2015. "The Value of Bosses." *Journal of Political Economy* 33(4): 823–61.

Marshall, Alfred. 1890. *Principles of Economics.* New York: Macmillan and Co.

Mas, Alexandre. 2006. "Pay, Reference Points, and Police Performance." *Quarterly Journal of*

Economics 121(3): 783–821.

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review* 99(1): 112–45.

Mobbs, Shawn, and Charu G. Raheja. 2011. "Internal Managerial Promotions: Insider Incentives and CEO Succession." *Journal of Corporate Finance* 18(5): 1337–53.

Moretti, Enrico, and Jeffrey M. Perloff. 2002. "Efficiency Wages, Deferred Payments, and Direct Incentives in Agriculture." *American Journal of Agricultural Economics* 84(4): 1144–55.

Nalebuff, Barry J., and Joseph E. Stiglitz. 1983. "Prizes and Incentives: Towards a General Theory of Compensation and Competition." *Bell Journal of Economics* 14(1): 21–43.

Neumark, David, and Wendy Stock. 1999. "Age Discrimination Laws and Labor Market Efficiency." *Journal of Political Economy* 107(5): 1081–1125.

O'Reilly, Charles A., III, Brian G. Main, and Graef S. Crystal. 1988. "CEO Compensation as Tournament and Social Comparison: A Tale of Two Theories." *Administrative Science Quarterly* 33(2): 257–74.

Rosen, Harvey S. 1976. "Taxes in a Labor Supply Model with Joint Wage-Hours Determination." *Econometrica* 44(3): 485–507.

Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy* 82(1): 34–55.

Ross, Stephen A. 1973. "The Economic Theory of Agency: The Principal's Problem." *American Economic Review* 63(2): 134–39.

Roy, Donald. 1952. "Quota Restriction and Goldbricking in a Machine Shop. *American Journal of Sociology* 57(5): 427–42.

Roy, Donald. 1954. "Efficiency and 'The Fix': Informal Intergroup Relations in a Piecework Machine Shop." *American Journal of Sociology* 60(3): 255–66.

Ruth, Jan-Erik, and James E. Birren. 1985. "Creativity in Adulthood and Old Age: Relations to Intelligence, Sex, and Mode of Testing." *International Journal of Behavioral Development* 8(1): 99–109.

Ryan, Joseph J., Jerome M. Sattler, and Shane J. Lopez. 2000. "Age Effects on Wechsler Adult Intelligence Scale-III Subtests." *Archives of Clinical Neuropsychology* 15(4): 311–17.

Shearer, Bruce. 2004. "Piece Rates, Fixed Wages, and Incentives: Evidence from a Field Experiment." *Review of Economic Studies* 71(2): 513–34.

Smith, Adam. 1776. *The Wealth of Nations.* Scotland: W. Strahan and T. Cadell.

Titmuss, Richard M. 1970. *The Gift Relationship: From Human Blood to Social Policy.* London: Unwin & Allen.

Weitzman, Martin L. 1984. *The Share Economy: Conquering Stagflation.* Cambridge, MA: Harvard University Press.

Williamson, Oliver E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications.* New York: Free Press.

Williamson, Oliver E. 1985. *The Economic Institution of Capitalism.* New York: Free Press.

Zinsmeister, Karl. 2016. *The Almanac of American Philanthropy.* Washington, DC; The Philanthropy Roundtable.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 215–238

Nonmonetary Incentives and the Implications of Work as a Source of Meaning

Lea Cassar and Stephan Meier

Economists typically assume in their models that work involves an exchange of time and effort for money. Work is considered "painful" because it requires "costly" effort and because of the opportunity cost of reduced leisure time. The economic implications of this view are straightforward. For incentive theory, the solution to any agency problem will be to design the optimal *monetary* incentive scheme. The role of human resource management boils down to offering the right monetary incentives. In addition, labor supply will be determined uniquely by the trade-off between utility from income and utility from leisure.

Nonmonetary job characteristics have received limited attention by economists when thinking about productivity and willingness to work. But the assumption that monetary compensation is what mainly matters for motivation at work is at odds with a number of observations. Close to home, what drives most academics to the university on a given day (including evenings and weekends) is not the money (otherwise we would work in the private sector) or the stability of the income stream (because the probability of losing a job is close to zero for a tenured academic). In fact, Stern (2004) shows that "scientists pay to be scientists." What motivates academics is the drive to contribute to our subject, applying our skills in solving intellectual challenges, the satisfaction of conducting our own research agendas, and what feels like an imperative to explain these ideas to others. Similarly, entrepreneurs often face

■ *Lea Cassar is Assistant Professor of Behavioral Managerial Economics, University of Cologne, Cologne, Germany. Stephan Meier is James P. Gorman Professor of Business Strategy, Columbia Business School, New York City, New York. Their email addresses are lcassar@uni-koeln.de and sm3087@gsb.columbia.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.215 doi=10.1257/jep.32.3.215

low risk-adjusted returns (Astebro, Herz, Nanda, and Weber 2014). One potential explanation for this phenomenon is that there are nonpecuniary aspects of being an entrepreneur, like a "utility premium" from being self-employed that seems to be related to the freedom to make autonomous decisions (Hamilton 2000; Benz and Frey 2008b).

Indeed, workers in many jobs act as if they care about more than just the highest paycheck. Consulting firms, like Great Places to Work, advise companies about creating motivating corporate cultures and create rankings about the best companies to work for based on comprehensive surveys about employees' attitudes toward their workplace. Such ratings abound, including the "Fortune 100 best companies to work for" and "Best Small & Medium Workplaces." Workers seem to care about more than income. In surveys of students and workers, 72 percent and 53 percent, respectively, say that "A job where I can make an impact" is very important or essential to their happiness (Net Impact 2012). Among chief executive officers, 59 percent think that "top talent prefers to work for organizations with social values which are aligned to their own" (PricewaterhouseCoopers 2016, p. 13).

In the past, there has been a disciplinary split in the attention paid to topics like nonmonetary compensation and benefits, workers' satisfaction, and intrinsic motivation to work. On one side, textbooks in human resources, management, or organizational behavior devote considerable attention to these issues; on the other side, these topics are covered only marginally, if at all, in textbooks about labor economics, contract theory, and organizational economics (for example, see Bolton and Dewatripont 2004; Gibbons and Roberts 2013; Cahuc and Zylberberg 2004).

But empirical research in economics has begun to explore the idea that workers care about nonmonetary aspects of work. In an early study, Ichniowski, Shaw, and Prennushi (1997) used data from steel production lines to show that a combination of incentive pay and flexible job assignment leads to substantial productivity increases compared to the more traditional practice of narrow job definition, strict working rules, and hourly pay with close supervision. Monetary incentives mattered, but were not the only motivator (for discussion in this journal, see Ichniowski and Shaw 2003). Since then, an increasing number of economic studies using survey and experimental methods have shown that nonmonetary incentives and nonpecuniary aspects of one's job have substantial impacts on job satisfaction, productivity, and labor supply. By drawing on this evidence and relating it to the literature in psychology, this paper argues that work represents much more than simply earning an income: for many people, work is a *source of meaning*.

In the next section, we give an economic interpretation of meaningful work and emphasize how it is affected by the mission of the organization and the extent to which job design fulfills the three psychological needs at the basis of self-determination theory: autonomy, competence, and relatedness (Deci and Ryan 1985, 2000; Ryan and Deci 2000). We point to the evidence that not everyone cares about having a meaningful job and discuss potential sources of this heterogeneity. We sketch a theoretical framework to start thinking about how to formalize work as a source of meaning and how to incorporate this idea into agency theory and labor supply models. We discuss how workers' search for meaning may affect the design of monetary and nonmonetary incentives. We conclude by suggesting some insights and open questions for future research.

Meaning in Work beyond Money

A long tradition in organizational behavior and organizational psychology argues that individuals get meaning from their work that extends beyond financial compensation (for a review, see Rosso, Dekas, and Wrzesniewski 2010). But in economics, there has been relatively little discussion about the desire for "meaning," although some notable exceptions include Loewenstein (1999), Karlsson, Loewenstein, and McCafferty (2004), and Chater and Loewenstein (2016).

A growing body of evidence suggests that nonmonetary factors are potentially important for motivation and productivity. As a starting point, Frank (1996) assembles some interesting facts: Cornell undergraduates report taking a 50 percent pay cut as an ad copywriter for the American Cancer Society compared to Camel Cigarettes; witnesses favoring cigarette regulation appear as volunteers while witnesses in the same lawsuits for the cigarette lobby have to be heavily compensated (even though the former are more qualified experts); and public interest lawyers accept much lower wages than associates in private law firms. According to Dur and Lent (2018), close to 77 percent of responders in their sample of 100,000 respondents across 47 countries from the International Social Survey Program Work Orientations Waves report that a job useful for society is important or very important for them. Bryce (2018) uses the American Time Use Survey and the UK Annual Population Survey and finds that work is reported to be more meaningful than consumer purchase, socializing, relaxing, or leisure. Jobs that are both high on personal autonomy and direct pro-social impact are rated as the most meaningful, including the jobs of health professionals, therapists, nurses, midwives, teachers, lecturers, and social workers.

Of course, survey evidence will always have difficulty in showing a causal connection between meaningful work and motivation or productivity, but evidence on this point is emerging, especially in the experimental literature. Ariely, Kamenica, and Prelec (2008) conduct an experiment in which performance and labor supply are affected by whether the job has some sort of point. The authors manipulate the meaningfulness of a task—specifically, assembling a Bionicle Lego—by varying whether the output resulting from subjects' work was destroyed immediately after completion, or whether it was kept intact. When the output of the task is destroyed, subjects have a 40 percent higher reservation wage to do the task compared to when the output is not destroyed. Chandler and Kapelner (2013) extend these results to a field experiment (see also Kosfeld, Neckermann, and Yang 2017). The authors hired M-Turk workers to label tumor cells, but some workers were explicitly told the purpose of their task was to help researchers identify tumor cells while other workers were not. When the task was framed in terms of meaning, workers were more likely to participate and, conditional on participating, they labeled a higher quantity of images.

What factors are likely to increase workers' feeling of meaning in their job? Here, we first discuss the role of the mission of the organization and then turn to three main aspects of job design.

The Role of the Mission

The mission of an organization, or the lack of a mission, can affect how employees perceive their own purpose. In general, an organization or a job with a social mission will be more likely to fulfill workers' drive for sense-making in their actions as part of a bigger social context, and for creating social bonds between the workers and the rest of the world. Both this act of sense-making and the connection to others are important ingredients of meaning (Karlsson, Loewenstein, and McCafferty 2004).

Evidence from nonprofit organizations shows that many people place a high value on working in a job with a pro-social mission and alter their work effort accordingly. For example, in a field experiment with fundraising callers, Grant (2008) shows that making the social purpose of the callers' job more salient increased the number of pledges earned by 124 percent and the amount of donations raised by 152 percent. In addition, workers appear willing to give up money to work in the nonprofit sector. Several studies suggest that nonprofit workers earn less than for-profit workers in comparable occupations (Preston 1988; Handy and Katz 1998; Leete 2001; Jones 2015). Furthermore, nonprofit workers are more likely to report a higher ideal number of hours worked (Lanfranchi, Narcy, and Larguem 2010).

While many organizations (or tasks within the organization) do not have an obvious and direct social purpose, they can nonetheless seek to create this meaning or purpose through "a concrete goal or objective for the firm that reaches beyond profit maximization" (Henderson and Van den Steen 2015, p. 327). For example, an investment bank might seek opportunities to create meaning through socially responsible business practice or by engaging in philanthropic activities. Indeed, an increasing number of companies are paying attention to corporate social responsibility, and not just companies with a long history of doing so, like Patagonia. For example, SABMiller, until recently the second-largest brewer in the world, invested heavily in reducing water usage in its beer production and in promoting sustainable water management where it operated (Mennel and Wong 2015). Many large firms like Deloitte or Bank of America are growing their pro bono volunteering programs (Novick O'Keefe 2016). Among the largest 250 companies in the world, 92 percent produced a corporate social responsibility report in 2015, up from 64 percent in 2005. Fortune Global 500 firms now spend around $15 billion a year on corporate social responsibility activities (Smith 2014), while US and European markets have over $2 trillion and €200 billion in certified socially responsible assets (Kitzmueller and Shimshack 2012).

These investments not only serve to make the firm's image more attractive to (socially responsible) costumers; sustainability initiatives, corporate foundations, employee volunteer programs, and donations to charity are also serving as tools for attracting and motivating employees. Firms such as IBM, General Motors, or Microsoft explicitly use their corporate social responsibility efforts to attract potential employees (Kitzmueller and Shimshack 2012). Other firms explicitly mention the impact of their work. For example, Medtronic says in the "career opportunities" section of their website (at http://europe.medtronic.com/xd-en/about/careers.html): "Careers that change lives: Do meaningful work, make a difference, and improve lives—starting with your own."

Such pro-social initiatives can increase effort, increase retention, and even lower employees' wage demands. For example, Burbano (2016) hired workers on two online marketplaces, provided them with either a message about the corporate social responsibility activity of the firm or with information about the work, and then elicited reservation wages. Being provided with information about an employer's social responsibility reduced reservation wages by 12 percent in one marketplace and 44 percent in the other. In a field experiment, Hedblom, Hickman, and List (2016) show that data-collecting jobs having a social mission (benefiting underprivileged children) increased the number of interested candidates by 26 percent. Moreover, the social mission component attracted more-productive workers with higher-quality work output who choose to work longer hours. Bode, Singh, and Rogan (2015) looked at management consultants engaged in social initiative projects, which involve the same tasks as commercial projects but are conducted with nonprofit organizations. Consultants accepted a lower wage while on those projects, and their probability of leaving the firm decreased by about 30 percent after participating in a social initiative program.

In general, social initiatives and mission increase job satisfaction. The earlier survey evidence suggested this connection, but Gosnell, List, and Metcalfe (2016) provide causal evidence from a field experiment with Virgin America pilots. They randomly offered charitable contributions for meeting fuel emission targets. While this treatment had the same effect on fuel efficiency as just providing the target (which is not surprising given that meeting the fuel emission target is already an environmental friendly action), job satisfaction increased by 6.5 percent.

Other studies also show that pro-social incentives in the form of charitable donations linked to work activity, where the financial reward is not paid to the worker but rather donated to a charity, increases the effort of workers both in the lab and in the field. In an online real-effort experiment, Tonin and Vlassopoulos (2015) finds that pro-social incentives lead to a 13 percent rise in productivity, regardless of their form (lump sum or related to performance) or strength. A positive effect of charitable donations on workers' effort is also found in Imas (2014); Charness, Cobo-Reyes, and Sánchez (2016), DellaVigna and Pope (2018), and Cassar (forthcoming). In particular, Imas (2014) and Charness et al. (2016) also show that when stakes are low, pro-social incentives lead to higher performance than standard incentives schemes.

Firms and business schools are noticing. The *2018 Deloitte Global Human Capital Trends* report summarizes: "Based on this year's global survey of more than 11,000 business and HR leaders, as well as interviews with executives from some of today's leading organizations, we believe that a fundamental change is underway. Organizations are no longer assessed based only on traditional metrics such as financial performance, or even the quality of their products or services. Rather, organizations today are increasingly judged on the basis of their relationships with their workers, their customers, and their communities, as well as their impact on society at large—transforming them from business enterprises into social enterprises." This trend is reflected in top business schools' curriculums. Between 2003 and 2009, the number of courses on "social entrepreneurship" (enterprises with an explicit social mission) at top US MBA programs increased by 110 percent (Beal 2017). Business schools increasingly recommend that companies align their corporate social responsibility strategy with their business interest (for example, Porter and Kramer 2007) or use it to motivate workers and win the war for talent (for example, Bhattacharya, Sen, and Korshun 2008). Economists have also started to devote their attention to social enterprises (Besley and Ghatak 2017).

The organizational mission, however, does not need to be charitable or "social" in a strict sense to increase meaning: it can represent any mission that is valuable to the workers (as modeled in the seminal paper by Besley and Ghatak 2005). In a recent field experiment, Carpenter and Gong (2016) shows that workers who stuffed letters to raise funds for political campaigns were 72 percent more productive if they worked for their favorite party rather than for the opposite party. People devote considerable amount of time and effort to open source initiatives, often anonymously and without financial returns, because they believe in the free diffusion of knowledge. For instance, Wikipedia is the world's sixth most popular website, comprising more than 35 million articles written by more than 55 million registered editors (Gallus 2016). Gartenberg, Prat, and Serafeim (2016) use a broader definition of meaning based in part on whether employees perceive that the management of a firm exhibits a high degree of clarity in setting goals. Firms with employees that perceive their job to be meaningful perform better financially (as measured by return on assets and Tobin's *q*) if the firm's management has a clear strategy.

Taken together, these findings and facts point to the importance of job mission as a source of meaning and, in turn, of intrinsic motivation to work.

The Role of Job Design

A growing body of evidence suggests that workers highly value certain nonmonetary job dimensions rooted in the three human psychological needs at the basis of self-determination theory (Deci and Ryan 1985, 2000; Ryan and Deci 2000): the needs for autonomy, competence, and relatedness.

Workers' autonomy in decision-making is an important determinant of job satisfaction, which in turn predicts economic behavior such as labor market mobility (Freeman 1978; Clark 2001) and productivity (Oswald, Proto, and Sgroi 2015). Freeman and Kleiner (2000) uses the Workplace Representation and Participation Survey, a nationally representative survey of nongovernment employees carried out in 1994–95 by Princeton Survey Research Associates, to show that workers in

"Employee Involvement" programs report themselves as very satisfied with "the influence they have in company decisions that affect their job or work life" compared to other workers. Similarly, Benz and Frey (2008a) use the nationally representative German Socio-Economic panel to show that people who work in smaller firms are more satisfied with their job and that this effect can be attributed to flatter hierarchies and higher levels of independence on the job. Using the same dataset, Bartling, Fehr, and Schmidt (2013) find that work autonomy and absence of monitoring is strongly associated with higher job satisfaction, even after controlling for several sociodemographics and occupational characteristics.

The literature on self-employment also shows that workers assign high value to autonomous decision-making. Using US data, Hamilton (2000) finds that people choose to become self-employed in spite of the low average return because self-employment offers nonmonetary benefits such as "being your own boss." Using data from 23 countries, Benz and Frey (2008b) show that the utility premium from self-employment is attributed to more interesting work and greater autonomy. In general, many people seem to value autonomy and flexibility. Chen, Chevalier, Rossi, and Oehlsen (2017) estimate for Uber drivers that their supplier surplus is about twice that of a less flexible arrangement. In a 2012 report (Gandia 2012), 64 percent list as main reason for becoming a freelancer the greater flexibility and freedom, while only 7.5 percent give a financial reason, such as higher income. Finally, experimental evidence indicates that reducing individuals' decisional autonomy negatively affects their work effort (Falk and Kosfeld 2006; Bartling, Fehr, and Schmidt 2012; Fehr, Herz, and Wilkening 2013).

It is then not surprising that many innovative companies seek to establish an organizational culture that favors workers' autonomy and entrepreneurial spirit. For Silicon Valley companies, promoting "entrepreneurship" rather than a "stewardship" culture among their employees is a prerequisite for success (Hamel 1999). Talented people do not need to be monitored. They need to be empowered.

A feeling of competence arises when workers are able to apply their talents, skills, and/or knowledge to achieve a certain goal. In fact, engaging in an activity that one is good at is generally pleasant (Loewenstein 1999). Personal and social recognition can play a substantial role in fostering a feeling of competence (as extensively discussed in Ellingsen and Johannesson 2007) and, in turn, of doing meaningful work.

Survey evidence shows that skill utilization is associated with job satisfaction and occupational choice. Using national US surveys, Eden (1975) and Hundley (2001) find that self-employed workers are more satisfied with their jobs than employees, because their work provides more skill utilization in addition to autonomy and flexibility. Using a nationally representative survey in Chile, Cassar (2010) shows that feeling competent and autonomous in one's job increases satisfaction and can explain the utility premium from self-employment. In addition, experimental studies show that social recognition in the form of nonmonetary awards increases workers' effort. In a field experiment in which students had to solve a data entry task, Kosfeld and Neckermann (2011) show that symbolic awards increase performance by 12 percent. Gallus (2016) shows in a field experiment involving Wikipedia that symbolic awards increase the share of new editors that remain active by 20 percent. Similar effects of awards on motivation are found in a number of studies (Gibbs, Neckermann, and Siemroth 2017; Ashraf, Bandiera, and Jack 2014; Chan, Frey, Gallus, and Torgler 2014). For a review of research on the effect of awards, see Gallus and Frey (2016).

Worker's feelings of relatedness and being connected with colleagues are another important aspect of meaning, which Karlsson, Loewenstein, and McCafferty (2004) refers to as "meaning as social extension of oneself." Social identity theory argues that workers who identify with the members and goals of their organization exert more effort, provide more public goods, coordinate their efforts better, and therefore, will be more productive (Akerlof and Kranton 2005, 2008). Empirical evidence shows that group identity increases public good provision (for example, Goette, Huffman, and Meier 2006) and facilitates coordination (for example, Chen and Chen 2011), and having positive social relations at work increases job satisfaction (Morgeson and Humphrey 2006). However, Bandiera, Barankay, and Rasul (2010) show that the effect of social incentives may depend on peers: in their setting, workers are more productive only when working with more able friends.

Whether workers feel connected to the organization and its members will obviously also depend on whether they feel treated fairly. A large body of evidence shows that workers want to be treated fairly—in particular, with regard to varying financial compensation across coworkers (Kahneman, Knetsch, and Thaler 1986; Bewley 1995; Fehr, Goette, and Zehnder 2009; Kaur 2014). In a month-long field experiment with Indian manufacturing workers, Breza, Kaur, and Shamdasani (2018) show that unequal pay that is perceived as unjustified has negative effects on both labor supply and work morale. As such, nepotism and favoritism within a company are likely to be detrimental for workers' productivity even beyond the standard monetary arguments. Nagin, Rebitzer, Sanders, and Taylor (2002) show that for workers who feel treated fairly, reducing monitoring did not lead to an increase in shirking and they summarize that "management's perceived empathy and fairness in dealing with employees may play an important role in reducing workplace opportunism" (p. 870).

Who Cares About Meaning?

Not all individuals search for meaning. Some do seem largely driven by financial motives. Many studies find heterogeneity in preferences about job attributes (for example, Wrzesniewski, McCauley, Rozin, and Schwartz 1997; Clark 2001; Bode, Singh, and Rogan 2015). Survey evidence from Net Impact (2012) *Talent Report* found that while 53 percent of workers consider making an impact as essential for their happiness, 47 percent do not seem to care as much. In a survey of science and engineering PhD candidates, Sauermann and Roach (2014) investigate how much they care about one nonpecuniary aspect of their future job—whether their future employer allows publishing. Slightly more than 20 percent do not care, although the median PhD candidate would be willing to give up 18 percent of base wage to be able to publish on the job. Fuchs-Schündeln (2009) finds that preferences for independence are heterogeneous across the population: for example, not all self-employed experience higher satisfaction in their job. In fact, self-employment can even decrease job satisfaction for "hierarchical types." In various lab and field experiments, the fraction of individuals who care about meaning and exhibit pro-social preferences varies from one-third to two-thirds (Fehr and Schmidt 1999; Fehrler and Kosfeld 2014; Tonin and Vlassopoulos 2015). The more recent evidence by Bruhin, Fehr, and Schunk (2016) actually suggests that other-regarding preferences are the rule rather than the exception.

The ultimate determinants of these variations in preferences remains an open question, but a number of important correlates have been discussed: age, income, ability, and other-regarding preferences.

Younger individuals tend to care more about meaning in their work (for example, Bode, Singh, and Rogan 2015; Clark 2001). One conjecture is that as people accumulate life experience, illusions about the meaning of work diminish and they are more likely to seek meaning in social relationships and other dimensions of life. However, it is extremely difficult to disentangle an explanation from possible omitted variables or from cohort effects. For example, financial responsibilities often rise with age, or preferences might shift with age, or different birth cohorts may have distinctive values.

One might conjecture that work meaning only matters for higher-income individuals—after basic needs are fulfilled (along the lines of Maslow's 1943 theory of human motivation). But the evidence for this thesis is unclear. For example, workers with higher incomes work more hours and are more likely to become workaholic (Hamermesh and Slemrod 2005). Perhaps higher pay provides more incentive to work, or perhaps this higher-income group has higher intrinsic pleasure from work (Freeman 2008), or perhaps high-paying occupations are more meaningful in feelings of autonomy, competence, and relatedness. As we discuss in the section below, perhaps monetary compensation and work meaning are complementary.

Some studies show that most-productive and more pro-social individuals are also more interested in meaningful job attributes (for example, Bode, Singh, and Rogan 2015; Hedblom, Hickman, and List 2016; Serra, Serneels, and Barr 2011) and/or reacted more strongly to the addition of meaning (for example, Burbano 2016). Independent of the underlying explanation for such an association, these correlations could point to the importance of offering a meaningful job in the war for high-performing talent—beyond just offering a competitive financial compensation package.

In sum, heterogeneity in preferences for meaning is substantial—not only in terms of whether individuals care at all about work meaning, but also in terms of what aspect of the job they value. While the correlates and causes of this heterogeneity need further study, the key point in the next section is that the existence of heterogeneity has implications for human relations strategy, in terms of sorting of workers and designing incentive and screening devices by firms.

Implications for Incentive Theory and Labor Supply

How does the meaning of work affect models of incentive theory and labor supply? To tackle this question, we sketch a conceptual framework that incorporates work meaning and that allows consideration of the relationships between nonmonetary incentives, monetary compensation, effort, and labor supply. To date, the literature on nonpecuniary motives for work has been quite fragmented, and we hope that this approach can provide a unifying and coherent framework.

A Utility Function with Work Meaning

In a standard neoclassical model of work motivation, a worker's utility function depends on two arguments: utility from income as a means to consumption, which is generated by a combination of financial compensation w and a level of effort e (which can be interpreted as hours and/or intensity of work), and utility from leisure, which, in the case of incentive theory, is typically modeled as the disutility or cost of effort. A worker chooses the effort level that equalizes its marginal benefits in generating income with its marginal costs in giving up leisure.

We propose to extend this basic neoclassical model to include work meaning as the third argument in the utility function, which then includes: the utility from income Y as a means to consumption, a cost C of providing effort (or foregoing utility from leisure), and the utility from meaningful work. Modeling the meaning of work as a separate term from effort cost is consistent with the interpretation of effort costs as preferences for leisure in a labor supply model and with previous theoretical work on intrinsic motivation (for example, Besley and Ghatak 2005; Delfgaauw and Dur 2007, 2008; Cassar 2016).[1] Here we focus on the case where work meaning enters positively into the utility function, so that work is a source of intrinsic motivation. However, a severe lack of work meaning (or "alienation") would enter negatively into the utility function. This utility function can be written as:

$$U = Y(w, e) + M(\boldsymbol{\theta}, \mathbf{x}, e) - C(e),$$

where M is the production function of meaningful work and $\mathbf{x}$ is a four-dimensional vector of the four aspects of job meaning (mission, autonomy, competence, and relatedness), which can be treated as exogenous or as endogenously chosen by the employer. The 4-dimensional vector $\boldsymbol{\theta}$ captures the weight assigned by the agent to each of these four aspects: thus, it captures the heterogeneity in preferences for meaning.

[1] Furthermore, this approach allows us to keep the standard assumption of increasing and convex effort costs and still predict a positive amount of effort even in the absence of monetary incentives or threat to be fired. This prediction seems more compatible with our work experience as academics, with the huge numbers of volunteers worldwide, and with the fact that many public organizations are still operating in spite of the near-impossibility of firing their employees.

Notice our assumption that M depends on the worker's effort e. The earlier evidence suggests that meaningful work translates not only into job satisfaction, but also into higher productivity and effort. Hence, we do not interpret the meaning of work as a stock variable, or a fixed characteristic, or as an additional constant in the utility function. Rather, we interpret *work meaning* as a flow, in the sense that it happens simultaneously with *work*. In other words, we see work with meaning as an intrinsic output that needs some ingredients to be generated—one of which must be that the agent is actually exerting effort to work. We assume that the meaningfulness of work increases with each of the four job dimensions, as well as in the agent's effort.

In this framework, the marginal effect of effort on producing meaningful work should be nondecreasing in each of the four dimensions of job meaning. This assumption matters: if my job is meaningless, no matter how much effort I will put in, it will not generate much meaning. If my job is meaningful, by working more I can also produce more meaning in this work. As an example, if my job has a strong "competence" dimension, in the sense that it allows me to apply my skills to solve challenges, then by working hard I can derive very high meaning from work. On the contrary, if my job has a weak "competence" dimension, such that I do a repetitive and unskilled task with little acknowledgment, then no matter how much effort I exert, I will not derive much meaning.

Depending on the application, some dimensions of job meaning may be affected by the wage and, therefore, M can also be a function of financial compensation w. For example, the feeling of relatedness in a job may depend on the wage level. Similarly, the sense of competence could be affected by the wage if wages are interpreted by workers as a signal or recognition of the worker's talent. If financial compensation affects both consumption and work meaning, it suggests the possibility of complementarities between financial pay and job meaning.

This framework links previous (otherwise disconnected) theoretical models of workers' intrinsic motivation (for a review, see Rebitzer and Taylor 2011). For example, Delfgaauw and Dur (2007, 2008) assume that workers vary in their intrinsic motivation to exert effort, but they do not model where this intrinsic motivation comes from. Compared to these models, we endogenize intrinsic motivation by making it dependent on the four dimensions of job meaning. Benabou and Tirole (2003) consider how performance incentives offered by an informed principal can adversely affect an agent's perception of ability to perform a task, and in turn intrinsic motivation. In our framework, this aspect of workers' utility would be captured by workers' preferences over the competence dimension of meaning. Besley and Ghatak (2005) study the effect that matching an organization's mission to the agents' preferred mission has on monetary incentives, hence it focuses on our first dimension of meaning. Prendergast (2007, 2008) assumes that bureaucrats care about the well-being of their clients, which can also be synthesized as bureaucrats having mission preferences. Akerlof and Kranton (2005, 2008) and Henderson and Van den Steen (2015) emphasize the role of corporate identity for the design of incentives, which in our framework would be captured by workers' preferences for relatedness. The behavioral contracting model with reciprocal agents by Englmaier and Leider (2012) can also be incorporated in our framework through workers' preferences for relatedness. Our notion of job design also captures some aspects of "corporate culture" (for a review of an economic analysis of corporate culture, see Hermalin 2012). However, our modeling of job meaning does not focus on what are often taken to be common ingredients of organizational culture, such as repeated interactions, shared information, and convergence toward common beliefs and preferences (Martinez, Beaulieu, Gibbons, Pronovost, and Wang 2015). Rather, it emphasizes how a culture with a strong focus on mission, autonomy, competence, and relatedness increases job meaning and thus can be highly valued by the employees.

A productive research agenda could be devoted to deepening our understanding of the four nonmonetary incentives and of the function that generates work meaning. For instance, are these different dimensions complements or substitutes? Armouti-Hansen, Cassar, and Dereky (2018) take a first step by studying contracting in a setting where agents can be motivated both by the social mission of their job and by reciprocity concerns towards the principal. One of their findings is that these two dimensions of job meaning are complementary in sustaining efficiency wages and in increasing effort. As another question, is the feeling of being competent more relevant for autonomous tasks or in teamwork? The answer is not obvious. One might argue that autonomous decision-making and feeling of competence go hand in hand, or one might argue that designing a job that gives a feeling of competence is particularly relevant in teamwork, because the sense of acknowledgment and recognition will motivate workers to continue doing their share even in settings where free-riding might be tempting.

When Should Firms Invest in Meaning?

In a framework that includes these dimensions of job meaning, decisions of both workers and firms become more complex. Workers may face new trade-offs between different dimensions of job meaning, while companies will have to form beliefs about workers' preferences and decide whether and how to invest in job meaning. In general, a firm's decision to invest in job meaning will depend on the cost of providing meaning, the relative marginal return on the effort of adding meaning versus increasing income, and the composition of the labor force.

Providing job meaning is likely to be costly for the firm in both time and financial terms. It can also restrict firms' choice sets in terms of incentive schemes and monitoring activities. Tournaments and employee rankings are likely to increase pay inequalities, which undermines the bonds between co-workers (Cassar and Klein 2017) and can have negative consequences in terms of lower effort and antisocial behavior (Harbring and Irlenbusch 2011; Goette, Huffman, Meier, and Sutter 2012). Monitoring can be perceived as a lack of trust and undermine the sense of relatedness. However, job meaning can increase firms' benefits from investing in training and technologies, the use of which could potentially help workers to build more competence. Workers' desire for impact can make it profitable for firms to give workers more autonomy (Dur and Glazer 2008).

Finally, job meaning will affect the design of compensation schemes. Given the utility function with job meaning described above, it is straightforward to show that it is optimal for a profit-maximizing employer to invest in job meaning and to offer a lower piece-rate compared to the benchmark case. Hence, consistent with previous work on intrinsic motivation (like Besley and Ghatak 2005), identity theory (Akerlof and Kranton 2005), and behavioral contracting with reciprocal agents (Englmaier and Leider 2012), job meaning is likely to flatten the optimal wage schedule, emphasizing the role of substitution between job meaning and performance-based monetary incentives.[2]

Given budget constraints, firms compare the marginal return of monetary incentives on effort to the marginal return of nonmonetary incentives. Several experiments have compared the relative impact of offering more money versus offering higher pro-social mission in the form of higher charitable donations. Tonin and Vlassopoulos (2015) and DellaVigna and Pope (2018) find that monetary incentives are more effective in boosting effort than charitable incentives, while Imas (2014) and Charness, Cobo-Reyes, and Sánchez (2016) find that the opposite is true for lower incentive levels. But all these studies find that workers' responses are insensitive to the size of the charitable incentives: that is, stakes matter less for charitable than for monetary incentives.[3] While this evidence is extremely preliminary and specific, it may suggest that firms that start with high pay and no social mission should find a greater payoff from investing in meaning—as long as their employees care to some extent about job meaning. Conversely, industries or occupations characterized by high pro-social mission and low pay—perhaps nonprofits or occupations like teachers and social workers—have relatively less to gain by increasing job meaning than by increasing pay.

Seeking to create meaning through a change in mission or job design may require some fundamental changes in the organization. In terms of mission, it is widely believed that greenwashing, or pretending that a product is environmentally friendlier than it actually is, backfires. Carlos and Lewis (2017) show empirically that some firms even decide *not* to publicize environmental certifications to avoid a risk of being perceived as hypocritical. In Cassar and Meier (2017), we collaborated with an Italian company to hire workers on M-Turk to create slogans for the company's products. In this setting, introducing pro-social incentives caused workers to react negatively by creating fewer slogans. Nonmonetary incentives, when used instrumentally, can be worse than ineffective.

Finally, firm decisions about adding meaning might not only be about increasing effort, but about attracting certain types of agents. However, targeting motivated agents may not be simple, especially if the agents' preference over job meaning are unobservable (for examples, see Delfgaauw and Dur 2007, 2008; Prendergast 2007; Cassar 2016). Delfgaauw and Dur (2008) show that if effort is unverifiable,

[2] The results may be different in a setting where higher piece-rates can serve the purpose of increasing the feeling of competence and, therefore, job meaning.

[3] See also the evidence and structural model in DellaVigna, List, Malmendier, and Rao (2016).

nonmotivated workers will find public sector jobs highly attractive and may crowd out motivated workers. If effort is verifiable, it becomes possible to screen the workers, which will affect how much job meaning a principal wants to offer. Cassar (2016) allow job meaning (in the form of the project mission) to be endogenous and show that an organization will invest more in job meaning in environments in which effort is noncontractible. On the contrary, unobservable heterogeneity in workers' preferences for meaningful work will lead employers to underinvest in job meaning relative to the socially optimal level. It will also induce them to offer separating contracts on the meaning dimension. Journalism offers an example of such a practice, in which journalists can be employed either as staff, and thus benefit from a higher pay but enjoy less project flexibility, or as freelance workers, who would typically earn less but have more freedom in choosing how to write their articles.

However, the extent to which firms will be able to screen for workers who care about job meaning will depend on the nature of the relationship between job meaning and financial incentives—namely, on whether meaning and monetary incentives are complements or substitutes. As we discuss in the next section, this question remains open.

Paying Less for a Meaningful Job?

In principle, job meaning could be either a substitute or a complement to monetary compensation, which in turn will influence whether people accept lower pay for a meaningful job, or whether job meaning and pay tend to rise together. The evidence on this point is mixed.

One source of evidence would be to look at nonprofit jobs, based on the assumption that they are more likely to offer meaningful work. However, the evidence on wage differentials for nonprofit jobs is mixed (for example, Mocan and Tekin 2003; Ruhm and Borkoski 2003; Preston 1988). Apparently, such wage differentials only exist for certain positions or certain industries, and it depends on the share of meaning-driven workers in the market (Leete 2001; Jones 2015). Moreover, comparisons between the nonprofit and the private sector will be biased if more productive employees self-select into the private sector.

There is also evidence that firms that invest in corporate social responsibility are able to offer lower wages (Nyborg and Zhang 2013). Again, such studies do not represent a random sample of firms and employees, but instead are subject to sorting and self-selection biases. Some experimental evidence confirms that it can be possible to offer lower wages for meaningful jobs and that sorting can matter. As mentioned before, Burbano (2016) shows that providing information about an employer's social responsibility reduced reservation wages. In a principal–agent laboratory experiment, Cassar (forthcoming) shows that agents exert more effort when effort generates a donation to a charity (pro-social mission) and that profit-maximizing principals take advantage of this intrinsic motivation by offering 20 percent lower monetary incentives. However, this study also finds that principals who care a lot about charitable giving offer higher piece-rates in the presence of, rather than in the absence of, the donation. This finding suggests that jobs with meaning do not necessarily pay less than jobs without meaning, and that the preferences of the employer matter too.

In general, if job meaning and financial compensation were substitutes, and informational asymmetries make it hard to learn about the meaningfulness of the job, then a low wage would also function as a signal about the job being meaningful (Benabou and Tirole 2003; Sliwka 2007). Additionally, if individuals care about having a meaningful job as a signal to others and/or to themselves that they are pro-social, then high monetary incentives may weaken this signal (Frey 1997; Bénabou and Tirole 2006; Ariely, Bracha, and Meier 2009). For a discussion of crowding-out of intrinsic motivation, see Gneezy, Meier, and Rey-Biel (2011) and Frey and Jegen (2001).

Evidence from experimental studies is mixed regarding whether lower incentives affect the pro-sociality of the applicant pool. Dal Bo, Finan, and Rossi (2013) and Ashraf, Bandiera, and Lee (2015) do not find that higher monetary incentives affect the pro-sociality of the applicant pool for government positions in Mexico and Zambia, respectively. However, Deseranno (forthcoming) shows that more-lucrative positions offered by a new health promoter nongovernment organization in Uganda were perceived as being less pro-social and discouraged agents with high pro-social preferences from applying. As the positions in the latter study were new and the ambiguity about the task higher, it is possible that the difference in information asymmetry created differences in the results between the studies. However, if lower wages cannot attract more motivated agents, then offering low compensation might not be possible—even for meaningful jobs.

On the other side, some studies point towards complementarities in human resources management practices (for a recent overview, see Englmaier and Schuessler 2016) and thus between job meaning and financial compensation. The Ichniowski, Shaw, and Prennushi (1997) study mentioned earlier has shown in an industrial context that the *combination* of incentive pay with a flexible job assignment increases productivity, which implies that for one context, at least, the complementarities between monetary and nonmonetary incentives are important. Bartling, Fehr, and Schmidt (2012) show in an experimental study that such complementarities can endogenously lead to two different types of jobs: "'bad' jobs with low discretion, low wages, and little rent-sharing, and 'good' jobs with high discretion, high wages, and substantial rent-sharing" (p. 834). The experiment shows the importance of screening for motivated agents and how competition for motivated workers leads employers to offer good jobs. In this experiment, "motivated" workers reacted negatively to low wages—even in the presence of high autonomy. Within this experimental setting, low wages cannot be offset by other meaningful job characteristics (in this case, offering a contract with full discretion) if they violate fairness norms.

How should the evidence about the meaning of work affect our thinking about monetary incentives in the workplace? Lazear (in this volume) reviews a wide range of evidence that workers care about financial compensation and that monetary incentives can increase effort. We view the evidence on job meaning as complementary with the evidence on monetary incentives in the workplace. For example, to the extent that job meaning and financial compensation are complements, our analysis gives reasons, beyond consumption, why employees may value higher wages: it shows recognition for a person's work.

Moreover, our theory offers predictions for when monetary incentives are likely to be effective. For instance, in a work environment deprived of job meaning—that is, the task is repetitive and boring, effort is not acknowledged, no control is given over the production process—monetary compensation will be the only remaining motivator. Unfortunately, there seems to be many jobs that provide little meaning. According to Gallup polls, more than 67 percent of US employees and more than 86 percent of employees worldwide report being not engaged in their jobs over the past 15 years (Mann and Harter 2016). Dur and Lent (2018) find that 8 percent of workers consider their job to be socially useless, while an additional 17 percent are doubtful on the issue. Most of these workers work in private-sector jobs involving simple and routine tasks, as well as jobs in finance, sales, marketing, and public relations. Technological advances might increase the number of workers who still face the problem of alienation in the modern workplace. *The Economist* magazine (2015) published an article called "Digital Taylorism: A modern version of 'scientific management' threatens to dehumanise the workplace."[4]

An interesting and open question is whether monetary incentives are likely to prove as effective in environments that are characterized by a high level of job meaning (including a sufficiently high wage that satisfies workers' need for both relatedness and competence). The answer depends on how much weight workers attribute to consumption relative to job meaning and how these terms enter into the utility functions. Perhaps if job meaning is already high, financial incentives can still be very effective in incentivizing people to work harder in order to increase their consumption. Alternatively, perhaps when the role of job meaning is already quite powerful in motivating effort, monetary incentives cannot really add much in this context. Either way, the answers will be very relevant for firms that want to win the war for talent.

How Meaning Affects Labor Supply

Preferences over work meaning also have implications for labor supply. One can apply the utility function described earlier to a model of labor supply by substituting the variable "effort" with "hours" and by replacing the cost of effort with the utility derived from leisure time. A worker maximizes this utility function by deciding how many hours to work. In a model that did not include work meaning,

[4] For other empirical evidence of alienation at work, Shantz, Alfes, and Truss (2014) use data from 227 employees in a manufacturing organization in the UK and find a strong positive association between lack of perceived job meaning, skill utilization, and decision power, on the one hand, and work alienation, on the other, which, in turn, is found to be positively correlated with emotional exhaustion and lower levels of well-being.

the marginal benefit of an additional hour worked is the wage. Hence, the wage would be set equal to the marginal rate of substitution between leisure and income. In a model with work meaning, the worker will gain additional meaning from more hours worked, which will raise the number of hours worked compared with the standard neoclassical optimality condition. In this setting, any reduction in employment will hit utility on two margins: lower income and lower meaning derived from work.

This perspective has policy implications. When government is thinking about the value to place on reducing unemployment, it should take into account both the loss of income and the loss of meaning caused by unemployment. In addition, workers who gain meaning from work will be less likely to reduce their hours than workers motivated by income alone. For this reason, work meaning makes labor supply (and thus employment) less procyclical than otherwise.

There is some evidence to support the connection from quantity of labor supplied to a sense of meaning. Research from happiness surveys has established the negative relationship between unemployment and well-being beyond income effects, and monetary unemployment benefits do not seem to fill this gap (for a literature review, see Frey and Stutzer 2010). Moreover, unemployment and retirement can be detrimental for health (for example, Kuhn, Wuellrich, and Zweimüller 2010; Fischer and Sousa-Poza 2009). However, we are still lacking empirical evidence that clearly shows that the channel through which unemployment affects well-being and health is in fact the reduction in meaning.

How should one reconcile a positive effect of job meaning on labor supply with the evidence that many people choose to take early retirement (Lazear 1986)? Again, we have to consider the heterogeneity in people's preferences over work meaning. Perhaps early retirement choices are mainly driven by those workers who derive little meaning from their jobs. In fact, research studies in health and sociology suggest that low job control—defined as "workers' authority to make decisions concerning their own activities and skill usage"—and bureaucratic workload (such as paperwork and meetings) are predictors of early retirement (Elovainio et al. 2005; Van Droogenbroeck and Spruyt 2014). Furthermore, even if workers with meaningful jobs chose to retire as soon as Social Security benefits begin, we do not know the counterfactual—namely, how early they would have chosen to retire if their work did not have any meaning.

Finally, notice that the connections from work meaning to labor supply sketched here make the big assumption that meaning is derived from work, and not from leisure. This assumption is in line with views that individuals get "Sunday neurosis," "that kind of depression which afflicts people who become aware of the lack of content of their lives when the rush of the busy week is over and the void within themselves becomes manifest" (Frankl 1959, p. 107). Of course, leisure activities can provide meaning, too. But it seems worth reconsidering the emphasis and interrelationships that economics has traditionally given to "work," "leisure," and "human goals." Perhaps the focus has been too narrow.

Conclusions and Directions for Future Research

Many workers care about more than financial compensation in their job. Nonmonetary incentives often matter, too. A firm's mission and the design of one's job can create meaning and purpose for employees. As a result, firms will have reason to care about meaning of work. We believe economists can usefully contribute to the debate about the implications of meaningful work. We are not arguing that financial compensation is unimportant. Lazear (in this volume) provides an excellent review of monetary incentives in certain organizations. But we believe that in order to manage modern organizations and understand the future of work, studying workers' nonmonetary motives will be crucial.

As the discussion in this paper has shown, there are a large number of open questions both theoretically and empirically about the effects and the limits of meaning as a nonmonetary incentive. Here are some aspects worth exploring.

First, although the framework of work meaning sketched here can serve as a starting point, a crucial step will be to develop formal models of work meaning. For example, we need to explore how the four dimensions of meaning interact with each other, whether they are substitutes and complements, and how monetary incentives are affected and affect the different dimensions. We have also argued that work meaning is valuable *per se*, but if some individuals care about meaning in order to signal a certain image to others or themselves, the implications can be quite different (Kosfeld, Neckermann, and Yang 2017). Additionally, workers might care about meaningful work because it serves as a signal of an unobservable characteristic, like a firm's trustworthiness. In a field experiment on eBay, Elfenbein, Fisman, and McManus (2012) show that tying charity donations to a product serves as a signal that the seller's product will be of high quality. If firms that emphasize meaningful work are sending signals by doing so, then even workers who react little or not at all to job meaning per se might find nonmonetary incentives to be relevant.

Second, we need to collect more detailed data on the four dimensions of job meaning proposed in this paper. Just focusing on (and trying to measure) GDP per capita and income inequality is too limited. Recently, economists have started measuring preferences for workplace attributes beyond income, such as flexibility, particularly in students or call center employees (Wiswall and Zafar 2017; Mas and Pallais 2017). This is a great start, but more systematic measures (both objective and subjective) of all different dimensions of work meaning for a representative population are needed.

Third, heterogeneity in workers' preferences for meaning is important. Because the most meaning-driven agents have been shown to be the most productive and the most likely to contribute to public goods within organization, nonmonetary incentives become an important tool in the war for talent. Competition for such employees, informational asymmetries, and the possibility of screening are likely to play roles and should be further investigated (Kosfeld and von Siemens 2011; Bartling, Fehr, and Schmidt 2012). Also, while we considered heterogeneous preferences for meaning as exogenous, preferences for meaning might also be endogenous—that is, created and reinforced (Ashraf and Bandiera 2017). This opens new ways of thinking about how organizations might be able to foster preferences for meaning or how past experiences and education can influence people's motivation to work.

Fourth, the tradeoffs between nonmonetary and monetary incentives need to be better understood. This means measuring the relative importance of different types of incentives, and also studying the interaction and substitutability between meaning, financial incentives, and other human resources/management practices. For example, an issue not addressed in this paper is how sustainable it is to motivate employees through work meaning. Some aspects of meaning, like the job design elements of autonomy, competence, and relatedness affect how work is done daily. In contrast, corporate social responsibility investments may affect workers who are choosing between two employers, but have less effect on daily productivity (or they might even have negative consequences as discussed in List and Momeni 2017). However, we have been struck by the evidence that even when meaning is artificially added, it can affect people's effort and labor supply (think of the Lego experiment by Ariely, Kamenica, and Prelec 2008), and we suspect that the positive results from such artificial additions of meaning represent a lower bound of the effects of meaning on people's motivation to work.

Finally, technological advances seem likely to affect the meaning of work. New technologies can increase job meaning in that they enable a more flexible organization of work (such as telecommuting), eliminate repetitive tasks, and help employees develop competences. But technological advance can also be detrimental for job meaning in that it can increase division of labor and monitoring. More broadly, if technology affects job meaning of various jobs differently, it might not only affect income inequality but inequality in work meaning. Economists can inform both firms and public policy by integrating work meaning as one essential ingredient in their theories of work motivation and by studying its implications for organizational economics, labor policies, and beyond.

References

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives* 19(1): 9–32.

Akerlof, George A., and Rachel E. Kranton. 2008. "Identity, Supervision, and Work Groups." *American Economic Review* 98(2): 212–17.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99(1): 544–55.

Ariely, Dan, Emir Kamenica, and Drazen Prelec. 2008. "Man's Search for Meaning: The Case of Legos." *Journal of Economic Behavior & Organization* 67 (3–4): 671–77.

Armouti-Hansen, Jesper, Lea Cassar, and Anna Dereky. 2018. "Efficiency Wages with Motivated Agents." Unpublished paper.

Ashraf, Nava, and Oriana Bandiera. 2017. "Altruistic Capital." *American Economic Review* 107(5): 70–75.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery." *Journal of Public Economics* 120: 1–17.

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee. 2015. "Do-Gooders and Go-Getters: Career Incentives, Selection and Performance in the Public Service Delivery." Available at: https://www.hbs.edu/faculty/Pages/item.aspx?num=46043.

Astebro, Thomas, Holger Herz, Ramana Nanda, and Roberto A. Weber. 2014. "Seeking the Roots of Entrepreneurship: Insights from Behavioral Economics." *Journal of Economic Perspectives* 28(3): 49–70.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2010. "Social Incentives in the Workplace." *Review of Economic Studies* 77(2): 417–58.

Bartling, Björn, Ernst Fehr, and Klaus M. Schmidt. 2012. "Screening, Competition, and Job Design: Economic Origins of Good Jobs." *American Economic Review* 102(2): 834–64.

Bartling, Björn, Ernst Fehr, and Klaus M. Schmidt. 2013. "Discretion, Productivity, and Work Satisfaction." *Journal of Institutional and Theoretical Economics* 169(1): 4–22.

Bhattacharya, Chitra B., Sankar Sen, and Daniel Korschun. 2008. "Using Corporate Social Responsibility to Win the War for Talent." *MIT Sloan Management Review* 49(2): 37–44.

Beal, Adam M. 2016 [2017]. "Funding the Future: Social Enterprise on the Rise." *Huffington Post.* https://www.huffingtonpost.com/adam-m-beal/funding-the-future-social_b_9825656.html. First published 5/3/2016; updated 5/3/2017.

Bénabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70(3): 489–520.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78.

Benz, Matthias, and Bruno S. Frey. 2008a. "Being Independent is a Great Thing: Subjective Evaluations of Self-Employment and Hierarchy." *Economica* 75(298): 362–83.

Benz, Matthias, and Bruno S. Frey. 2008b. "The Value of Doing What You Like: Evidence from the Self-Employed in 23 Countries." *Journal of Economic Behavior & Organization* 68 (3–4): 445–55.

Besley, Timothy, and Maitreesh Ghatak. 2005. "Competition and Incentives with Motivated Agents." *American Economic Review* 95 (3): 616–36.

Besley, Timothy, and Maitreesh Ghatak. 2017. "Profit with Purpose? A Theory of Social Enterprise." *American Economic Journal: Economic Policy* 9(3): 19–58.

Bewley, Truman F. 1995. "A Depressed Labor Market as Explained by Participants." *American Economic Review* 85(2): 250–54.

Bode, Christiane, Jasjit Singh, and Michelle Rogan. 2015. "Corporate Social Initiatives and Employee Retention." *Organization Science* 26(6): 1702–20.

Bolton, Patrick, and Mathias Dewatripont. 2004. *Contract Theory.* MIT Press.

Breza, Emily, Supreet Kaur, and Yogita Shamdasani. 2018. "The Morale Effects of Pay Inequality." *Quarterly Journal of Economics* 133(2): 611–63.

Bruhin, Adrian, Ernst Fehr, and Daniel Schunk. 2016. "The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences." CESifo Working Paper 5744.

Bryce, Andrew. 2018. "Finding Meaning through Work: Eudaimonic Well-Being and Job Type in the US and UK." SERPS no. 2018004, Sheffield Economic Research Paper Series, May. https://www.sheffield.ac.uk/economics/research/serps/articles/2018004.

Burbano, Vanessa C. 2016. "Social Responsibility Messages and Worker Wage Requirements: Field Experimental Evidence from Online Labor Marketplaces." *Organization Science* 27(4): 1010–28.

Cahuc, Pierre, and André Zylberberg. 2004. *Labor Economics.* MIT Press.

Carlos, W. Chad, and Ben W. Lewis. 2017. "Strategic Silence: Withholding Certification Status as a Hypocrisy Avoidance Tactic." *Administrative Science*

Quarterly 63(1): 130–69.

Carpenter, Jeffrey, and Erick Gong. 2016. "Motivating Agents: How Much Does the Mission Matter?" *Journal of Labor Economics* 34(1): 211–36.

Cassar, Lea. 2010. "Revisiting Informality: Evidence from Employment Characteristics and Job Satisfaction in Chile." OPHI Working Paper no. 41, Oxford Poverty & Human Development Initiative.

Cassar, Lea. Forthcoming. "Job Mission as a Substitute for Monetary Incentives: Benefits and Limits." *Management Science.*

Cassar, Lea. 2016. "Optimal Contracting with Endogenous Project Mission." CESifo Working Paper 6181.

Cassar, Lea, and Arnd H. Klein. 2017. "A Matter of Perspective: How Experience Shapes Preferences for Redistribution." CESifo Working Paper 6302.

Cassar, Lea, and Stephan Meier. 2017. "Intentions for Doing Good Matter for Doing Well: The (Negative) Signaling Value of Prosocial Incentives." NBER Working Paper 24109.

Chan, Ho Fai, Bruno S. Frey, Jana Gallus, and Benno Torgler. 2014. "Academic Honors and Performance." *Labour Economics* 31: 188–204.

Chandler, Dana, and Adam Kapelner. 2013. "Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets." *Journal of Economic Behavior & Organization* 90 : 123–33.

Charness, Gary, Ramón Cobo-Reyes, and Ángela Sánchez. 2016. "The Effect of Charitable Giving on Workers' Performance: Experimental Evidence." *Journal of Economic Behavior & Organization* 131(Part A): 61–74.

Chater, Nick, and George Loewenstein. 2016. "The Under-Appreciated Drive for Sense-Making." *Journal of Economic Behavior & Organization* 126(Part B): 137–54.

Chen, M. Keith, Judith A. Chevalier, Peter E. Rossi, and Emily Oehlsen. 2017. "The Value of Flexible Work: Evidence from Uber Drivers." NBER Working Paper 23296.

Chen, Roy, and Yan Chen. 2011. "The Potential of Social Identity for Equilibrium Selection." *American Economic Review* 101(6): 2562–89.

Clark, Andrew E. 2001. "What Really Matters in a Job? Hedonic Measurement Using Quit Data." *Labour Economics* 8 (2): 223–42.

Dal Bó, Ernesto, Frederico Finan, and Martín A. Rossi. 2013. "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service." *Quarterly Journal of Economics* 128(3): 1169–1218.

Deci, Edward L., and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior.* Springer Science & Business Media.

Deci, Edward L., and Richard M. Ryan. 2000. "The 'What' and 'Why' of Goal Pursuits: Human Needs and the Self-Determination of Behavior." *Psychological Inquiry* 11(4): 227–68.

Delfgaauw, Josse, and Robert Dur. 2007. "Signaling and Screening of Workers' Motivation." *Journal of Economic Behavior & Organization* 62(4): 605–24.

Delfgaauw, Josse, and Robert Dur. 2008. "Incentives and Workers' Motivation in the Public Sector." *Economic Journal* 118 (525): 171–91.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao. 2016. "Estimating Social Preferences and Gift Exchange at Work." NBER Working Paper 22043.

DellaVigna, Stefano, and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Stuides* 85(2): 1029–69.

Deloitte. 2018. *2018 Deloitte Global Human Capital Trends: The Rise of the Social Enterprise.* Deloitte Insights. https://www2.deloitte.com/content/dam/insights/us/articles/HCTrends2018/2018-HCtrends_Rise-of-the-social-enterprise.pdf.

Deseranno, Erika. Forthcoming. "Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda." *American Economic Journal: Applied Economics.*

Dur, Robert, and Amihai Glazer. 2008. "The Desire for Impact." *Journal of Economic Psychology* 29(3): 285–300.

Dur, Robert, and Max van Lent. 2018. "Socially Useless Jobs." Tinbergen Institute Discussion Paper TI 2018-034/VII, https://papers.tinbergen.nl/18034.pdf.

Economist, The. 2015. "Digital Taylorism." September 10. https://www.economist.com/business/2015/09/10/digital-taylorism.

Eden, Dov. 1975. "Organizational Membership vs Self-employment: Another Blow to the American Dream." *Organizational Behavior and Human Performance* 13 (1): 79–94.

Elfenbein, Daniel W., Raymond Fisman, and Brian McManus. 2012. "Charity as a Substitute for Reputation: Evidence from an Online Marketplace." *Review of Economic Studies* 79(4): 1441–68.

Ellingsen, Tore, and Magnus Johannesson. 2007. "Paying Respect." *Journal of Economic Perspectives* 21(4): 135–50.

Elovainio, Marko, Pauli Forma, Mika Kivimäki, Timo Snervo, Risto Sutinen, and Marjukka Laine. 2005. "Job Demands and Job Control as Correlates of Early Retirement Thoughts in Finnish Social and Health Care Employees." *Work & Stress* 19(1): 84–92.

Englmaier, Florian, and Stephen Leider. 2012. "Contractual and Organizational Structure with Reciprocal Agents." *American Economic Journal:*

Microeconomics 4(2): 146–83.

Englmaier, Florian, and Katharian Schuessler. 2016. "Complementarities of Human-Resource Management Practices: A Case for a Behavioral-Economics Perspective." *Journal of Institutional and Theoretical Economics* 172(2): 312–41.

Falk, Armin, and Michael Kosfeld. 2006. "The Hidden Costs of Control." *American Economic Review* 96(5): 1611–30.

Fehr, Ernst, Lorenz Goette, and Christian Zehnder. 2009. "A Behavioral Account of the Labor Market: The Role of Fairness Concerns. *Annual Review of Economics* 1: 355–84.

Fehr, Ernst, Holger Herz, and Tom Wilkening. 2013. "The Lure of Authority: Motivation and Incentive Effects of Power." *American Economic Review* 103(4): 1325–59.

Fehr, Ernst, and Kaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114 (3): 817–68.

Fehrler, Sebastian, and Michael Kosfeld. 2014. "Pro-Social Missions and Worker Motivation: An Experimental Study." *Journal of Economics Behavior and Organization* 100: 99–110.

Fischer, Justina A. V., and Alfonso Sousa-Poza. 2009. "Does Job Satisfaction Improve the Health of Workers? New Evidence Using Panel Data and Objective Measures of Health." *Health Economics* 18(1): 71–89.

Frank, Robert H. 1996. "What Price the Moral High Ground?" *Southern Economic Journal* 63(1): 1–17.

Frankl, Victor E. 1959. *Man's Search for Meaning.* Beacon Press.

Freeman, Richard B. 1978. "Job Satisfaction as an Economic Variable." *American Economic Review* 68(2): 135–41.

Freeman, Richard B. 2008. "Why Do We Work More than Keynes Expected?" Chap. 9 in *Revisiting Keynes: Economic Possibilities for Our Grandchildren*, edited by Lorenzo Pecchi and Gustavo Piga. MIT Press.

Freeman, Richard B., and Morris M. Kleiner. 2000. "Who Benefits Most from Employee Involvement: Firms or Workers? *American Economic Review* 90(2): 219–23.

Frey, Bruno S. 1997. *Not Just for the Money: An Economic Theory of Intrinsic Motivation.* Edward Elgar.

Frey, Bruno S., and Alois Stutzer. 2010. *Happiness and Economics: How the Economy and Institutions Affect Human Well-Being.* Princeton University Press.

Frey, Bruno S., and Reto Jegen. 2001. "Motivation Crowding Theory." *Journal of Economic Surveys* 15(5): 589–611.

Fuchs-Schündeln, Nicola. 2009. "On Preferences for Being Self-Employed." *Journal of Economic Behavior & Organization* 71(2): 162–71.

Gallus, Jana. 2016. "Fostering Public Good Contributions with Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia." *Management Science* 63(12): 3999–4015.

Gallus, Janus, and Bruno S. Frey. 2016. "Awards: A Strategic Management Perspective." *Strategic Management Journal* 37(8): 1699–1714.

Gartenberg, Claudine, Andrea Prat, and George Serafeim. 2016. "Corporate Purpose and Financial Performance." Working Paper, Harvard Business School. https://hbswk.hbs.edu/item/corporate-purpose-and-financial-performance.

Gandia, Ed. 2012. *2012 Freelance Industry Report: Data and Analysis of Freelancer Demographics, Earnings, Habits and Attitudes.* August. https://s3.amazonaws.com/ifdconference/2012report/FreelanceIndustryReport2012.pdf.

Gibbons, Robert, and John Roberts, eds. 2013. *The Handbook of Organizational Economics.* Princeton University Press.

Gibbs, Michael, Susanne Neckermann, and Christoph Siemroth. 2017. "A Field Experiment in Motivating Employee Ideas." *Review of Economics and Statistics* 99(4): 577–90.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25(4): 191–210.

Goette, Lorenz, David Huffman, and Stephan Meier. 2006. "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence using Random Assignment to Real Social Groups." *American Economic Review* 96(2): 212–16.

Goette, Lorenz, David Huffman, Stephan Meier, and Matthias Sutter. 2012. "Competition between Organizational Groups: Its Impact on Altruistic and Antisocial Motivations." *Management Science* 58(5): 948–60.

Gosnell, Greer K., John A. List, and Robert Metcalfe. 2016. "A New Approach to an Age-Old Problem: Solving Externalities by Incenting Workers Directly." NBER Working Paper 22316.

Grant, Adam M. 2008. "The Significance of Task Significance: Job Performance Effects, Relational Mechanisms, and Boundary Conditions." *Journal of Applied Psychology* 93(1): 108–24.

Harbring, Christine, and Bernd Irlenbusch. 2011. "Sabotage in Tournaments: Evidence from a Laboratory Experiment." *Management Science* 57(4): 611–27.

Hamel, Gary. 1999. "Bringing Silicon Valley Inside." *Harvard Business Review* 77(5): 70–84,183.

Hamermesh, Daniel S., and Joel Slemrod. 2005. "The Economics of Workaholism: We Should Not Have Worked on This Paper." NBER Working

Paper 11566.

Hamilton, Barton H. 2000. "Does Entrepreneurship Pay? An Empirical Analysis of the Returns to Self-Employment." *Journal of Political Economy* 108(3): 604–31.

Handy, Femida, and Eliakim Katz. 1998. "The Wage Differential between Nonprofit Institutions and Corporations: Getting More by Paying Less?" *Journal of Comparative Economics* 26(2): 246–61.

Hedblom, Daniel, Brent R. Hickman, and John A. List. 2016. "Toward an Understanding of Corporate Social Responsibility: Theory and Field Experimental Evidence." Unpublished paper.

Henderson, Rebecca, and Eric Van den Steen. 2015. "Why Do Firms Have 'Purpose'? The Firm's Role as a Carrier of Identity and Reputation." *American Economic Review* 105(5): 326–30.

Hermalin, Benjamin E. 2012. "Leadership and Corporate Culture." Chap. 11 in *The Handbook of Organizational Economics*, edited by Robert Gibbons and John Roberts. Princeton University Press.

Hundley, Greg. 2001. "Why and When Are the Self-Employed More Satisfied with their Work?" *Industrial Relations: A Journal of Economy and Society* 40(2): 293–316.

Ichniowski, Casey, Kathryn Shaw, and Giovanna Prennushi. 1997. "The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines." *American Economic Review* 87(3): 291–313.

Ichniowski, Casey, and Kathryn Shaw. 2003. "Beyond Incentive Pay: Insiders' Estimates of the Value of Complementary Human Resource Management Practices." *Journal of Economic Perspectives* 17(1): 155–80.

Imas, Alex. 2014. "Working for the Warm Glow: On the Benefits and Limits of Prosocial Incentives." *Journal of Public Economics* 114: 14–18.

Jones, Daniel B. 2015. "The Supply and Demand of Motivated Labor: When Should We Expect to See Nonprofit Wage Gaps? *Labour Economics* 32: 1–14.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review* 76(4): 728–41.

Karlsson, Niklas, George Loewenstein, and Jane McCafferty. 2004. "The Economics of Meaning." *Nordic Journal of Political Economy* 30: 61–75.

Kaur, Supreet. 2014. "Nominal Wage Rigidity in Village Labor Markets." NBER Working Paper 20770.

Kitzmueller, Markus, and Jay Shimshack. 2012. "Economic Perspectives on Corporate Social Responsibility." *Journal of Economic Literature* 50(1): 51–84.

Kosfeld, Michael, and Susanne Neckermann. 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics* 3(3): 86–99.

Kosfeld, Michael, Susanne Neckermann, and Xiolan Yang. 2017. "The Effects of Financial and Recognition Incentives across Work Contexts: The Role of Meaning." *Economic Inquiry* 55(1): 237–47.

Kosfeld, Michael, and Ferdinand A. von Siemens. 2011. "Competition, Cooperation, and Corporate Culture." *RAND Journal of Economics* 42(1): 23–43.

Kuhn, Andreas, Jean-Phillipe Wuellrich, and Josef Zweimüller. 2010. "Fatal Attraction? Access to Early Retirement and Mortality." Available at SSRN: https://ssrn.com/abstract=1672168.

Lanfranchi, Joseph, Mathieu Narcy, and Makram Larguem. 2010. "Shedding New Light on Intrinsic Motivation to Work: Evidence from a Discrete Choice Experiment. *Kyklos* 63(1): 75–93.

Lazear, Edward P. 1986. "Retirement from the Labor Force." Chap 5. in *Handbook of Labor Economics*, vol. 1, edited by Orley C. Ashenfelter and Richard Layard. Elsevier.

Leete, Laura. 2001. "Whither the Nonprofit Wage Differential? Estimates from the 1990 Census." *Journal of Labor Economics* 19(1): 136–70.

List, John A., and Fatemeh Momeni. 2017. "When Corporate Social Responsibility Backfires: Theory and Evidence from a Natural Field Experiment." NBER Working Paper 24169.

Loewenstein, George. 1999. "Because It Is There: The Challenge of Mountaineering for Utility Theory." *Kyklos* 52(3): 315–43.

Mann, Annamar, and Jim Harter. 2016. "The Worldwide Employee Engagement Crisis." *Business Journal*, January. 7. https://news.gallup.com/businessjournal/188033/worldwide-employee-engagement-crisis.aspx.

Mas, Alexandre, and Amanda Pallais. 2017. "Valuing Alternative Work Arrangements." *American Economic Review* 107(12): 3722–59.

Maslow, A. H. 1943. "A Theory of Human Motivtion." *Psychological Review* 50(4): 370–96.

Martinez, Elizabeth A., Nancy Beaulieu, Robert Gibbons, Peter Pronovost, and Thomas Wang. 2015. "Organizational Culture and Performance." *American Economic Review* 105(5): 331–35.

Mennel, John, and Nate Wong. 2015. "Driving Corporate Growth through Social Impact: Four Corporate Archetypes to Maximize your Social Impact." *Deloitte Consulting Report.*

Mocan, Naci H., and Erdal Tekin. 2003. "Nonprofit Sector and Part-Time Work: An Analysis of Employer-Employee Matched Data on Child Care Workers." *Review of Economics and Statistics* 85(1): 38–50.

Morgeson, Frederick P., and Stephen E.

Humphrey. 2006. "The Work Design Questionnaire (WDQ): Developing and Validating a Comprehensive Measure for Assessing Job Design and the Nature of Work." *Journal of Applied Psychology* 91(6): 1321–39.

Nagin, Daniel S., James B. Rebitzer, Seth Sanders, and Lowell J. Taylor. 2002. "Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment." *American Economic Review* 92(4): 850–73.

Net Impact. 2012. *Talent Report: What Workers Want in 2012.* Net Impact.

Novick O'Keefe, Linda. 2016. "CSR Grows in 2016 as Companies Embrace Employees' Values." *Huffington Post, The Blog,* December 15.

Nyborg, Karine, and Tao Zhang. 2013. "Is Corporate Social Responsibility Associated with Lower Wages?" *Environmental and Resource Economics* 55 (1): 107–17.

Oswald, Andrew J., Eugenio Proto, and Daniel Sgroi. 2015. "Happiness and Productivity." *Journal of Labor Economics* 33(4): 789–822.

Porter, Michael E., and Mark R. Kramer. 2007. "Strategy and Society: The Link Between Competitive Advantage and Corporate Social Responsibility." *Harvard Business Review,* December.

Prendergast, Canice. 2007. "The Motivation and Bias of Bureaucrats." *American Economic Review* 97(1): 180–96.

Prendergast, Canice. 2008. "Intrinsic Motivation and Incentives." *American Economic Review* 98(2): 201–05.

Preston, Anne E. 1988. "The Effects of Property Rights on Labor Costs of Nonprofit Firms: An Application to the Day Care Industry." *Journal of Industrial Economics* 36(3): 337–50.

PricewaterhouseCoopers. 2016. "Redefining Business Success in a Changing World: CEO Survey." Technical report, PricewaterhouseCoopers 19th Annual Global CEO Survey.

Rebitzer, James B., and Lowel J. Taylor. 2011. "Extrinsic Rewards and Intrinsic Motives: Standard and Behavioral Approaches to Agency and Labor Markets." Chap. 8 in *Handbook of Labor Economics* vol. 4A, edited by Orley Ashenfelter and David Card. Elsevier.

Rosenberg, Morris. 1957. *Occupations and Values.* With the assistance of Edward A. Suchman and Rose K. Goldsen. Glencoe, IL: The Free Press.

Rosso, Brent D., Kathryn H. Dekas, and Amy Wrzesniewski. 2010. "On the Meaning of Work: A Theoretical Integration and Review." *Research in Organizational Behavior* 30: 91–127.

Ruhm, Christopher J., and Carey Borkoski. 2003. "Compensation in the Nonprofit Sector." *Journal of Human Resources* 38(4): 992–1021.

Ryan, Richard M., and Edward L. Deci. 2000. "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being." *American Psychologist* 55(1): 68–78.

Sauermann, Henry, and Michael Roach. 2014. "Not All Scientists Pay to be Scientists: PhDs' Preferences for Publishing in Industrial Employment." *Research Policy* 43(1): 32–47.

Serra, Danila, Pieter Serneels, and Abigail Barr. 2011. "Intrinsic Motivations and the Non-Profit Health Sector: Evidence from Ethiopia." *Personality and Individual Differences* 51(3): 309–14.

Shantz, Amanda, Kerstin Alfes, and Cathrine Truss. 2014. "Alienation from Work: Marxist Ideologies and Twenty-First-Century Practice." *International Journal of Human Resource Management* 25(18): 2529–50.

Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review* 97(3): 999–1012.

Smith, Alison. 2014. "Fortune 500 Companies Spend More than $15bn on Corporate Responsibility." *Financial Times,* October 12.

Stern, Scott. 2004. "Do Scientists Pay to Be Scientists?" *Management Science* 50(6): 835–53.

Tonin, Mirco, and Michael Vlassopoulos. 2015. "Corporate Philanthropy and Productivity: Evidence from an Online Real Effort Experiment." *Management Science* 61(8): 1795–1811.

Van Droogenbroeck, Filip, and Bram Spruyt. 2014. "To Stop or Not to Stop: An Empirical Assessment of the Determinants of Early Retirement among Active and Retired Senior Teachers." *Research on Aging* 36(6): 753–77.

Wiswall, Matthew, and Basit Zafar. 2017. "Preference for the Workplace, Investment in Human Capital, and Gender." *Quarterly Journal of Economics* 133(1): 457–507.

Wrzesniewski, Amy, Clark McCauley, Paul Rozin, and Barry Schwartz. 1997. "Jobs, Careers, and Callings: People's Relations to Their Work." *Journal of Research in Personality* 31(1): 21–33.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 239–258

The Changing (Dis-)Utility of Work

Greg Kaplan and Sam Schulhofer-Wohl

Most economists view monetary rewards as by far the most important aspect of jobs and careers. The disutility of supplying one hour of labor is assumed to be the same whether that hour is spent building cars on an assembly line, waiting tables at a restaurant, teaching a class, or pitching for the Chicago White Sox. In consequence, in conventional models, the tradeoffs workers make between consumption and leisure can be assessed solely by looking at hours worked and wages.

Yet it is obvious to many workers that a job involves more than just forfeiting some leisure time in return for a wage (Schwartz 2015). Jobs differ in the physical and mental toll they take on workers, as well as in the psychological rewards they provide, such as autonomy and meaning (Kalleberg 2011).

Our study asks how the major occupational shifts in the postwar period have manifested in changes in the nonpecuniary costs and benefits of work. Many fewer people work on assembly lines now than in 1950, while many more work in services and sales. Women and minorities have moved in large numbers into jobs where they once faced substantial barriers to entry. How have these shifts changed the aggregate amount of hardship or disutility that people experience from their work, the aggregate psychological rewards or utility that they derive from it, and the distribution of disutility and utility across the population?

■ *Greg Kaplan is Professor of Economics, University of Chicago, Chicago, Illinois, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Sam Schulhofer-Wohl is Senior Economist and Research Advisor, Federal Reserve Bank of Chicago, Chicago, Illinois. Their email addresses are gkaplan@uchicago.edu and samuel.schulhofer-wohl@chi.frb.org.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.239 doi=10.1257/jep.32.3.239

Drawing on data from the American Time Use Survey (ATUS), we examine six dimensions of workers' feelings about the time they spend on the job in different occupations: how happy, sad, and tired they are; how much stress and pain they experience; and how meaningful they find their work. We then calculate how economy-wide *average* feelings about work depend on the mix of occupations in the economy.

For some of the dimensions we study, such as pain, introspection offers an easy answer to the direction of change for most workers in developed countries. Although even today many jobs are undoubtedly physically demanding, we can easily compare how someone feels after a day of office work with how a farmer in the early 20th century must have felt after a day in the fields. Our results confirm that, in the aggregate, work has become less painful and less tiring in the postwar period.

But for other dimensions, where introspection does not provide easy answers, our study offers tentative new insights on the directions of change for US workers. We find substantial heterogeneity in how the nonphysical costs and benefits of work have changed over time. For women, the nonphysical aspects of work have on average become more positive over time: Women have shifted toward occupations that produce more happiness and meaningfulness and less sadness, while experiencing no change in stress. The story for men is more negative. Although they have shared in the reduction in pain and tiredness, they also have shifted toward occupations that produce more stress, less happiness, and less meaningfulness. The improvements for women and the reduction in meaningfulness for men appear to be concentrated among people at lower education levels. All this is not to deny, of course, that many workers even today have jobs that are painful, tiring, meaningless, saddening, or stressful—only that the share of such jobs is lower than in the past.

Our analysis of the nonpecuniary implications of changes in the occupational structure complements the large existing literatures on the wage and employment implications of these changes. For example, Herrendorf, Rogerson, and Valentinyi (2014) examine the aggregate employment and consumption effects of sectoral shifts from agriculture to manufacturing and then to services over the last two centuries. More recent changes in the occupational structure have been characterized by polarization—meaning the simultaneous growth of high-wage, high-skill jobs and low-wage, low-skill jobs, even as employment shrinks in the middle of the wage and skill distribution (Autor, Katz, and Kearney 2006). Studies of polarization have highlighted how these different patterns of job growth relate to the changing nature of tasks required by employers and the skills required to do different tasks (Acemoglu and Autor 2011). Although these analyses usually focus on the monetary returns to different types of skills, it is increasingly accepted that the nonmonetary returns to skill have also changed and that these changes differ sharply in the cross-section (Hamermesh 2001).

There is extensive debate over the macroeconomic forces that have led to polarization, but much less work on the consequences of polarization for workers' well-being, both individually and in the aggregate. With the recent rise in long-term unemployment in the United States and the rise in deaths related to substance abuse and suicide in the same time period (Case and Deaton 2017), it is becoming more evident that the loss of psychological benefits of work may be an important component of the overall costs of changes in employment.

One potential psychological benefit of work is its meaning (if any). The opening sentences of Ariely, Kamenica, and Prelec (2008) summarize the gap between the importance that workers place on meaning and the importance that economists place on it:

> Most children think of their potential future occupations in terms of what they will *be* (firemen, doctors, etc.), not merely what they will *do* for a living. Many adults also think of their job as an integral part of their identity. At least in the United States, "What do you do?" has become as common a component of an introduction as the anachronistic "How do you do?" once was, yet identity, pride, and meaning are all left out from standard models of labor supply.

That paper demonstrates the importance of meaning for workers' productivity in a laboratory setting. But there are also many examples of the strength of meaning as a motivating tool in real workplaces (Grant 2007, 2012). Against the backdrop of these micro-level examinations of the role of meaning in particular work environments, our study offers a macro perspective on the aggregate meaningfulness of work across the US economy.

On the cost side, economists typically think only of the opportunity cost of the time spent at work. But there are many features that make some jobs less desirable than others (Katz and Krueger 2016). For example, work can be so physically demanding that it leaves workers tired, injured or ill, or even kills them. Work has become dramatically less deadly over time (Aldrich 1997; Centers for Disease Control and Prevention 1999), perhaps as a result of occupational safety regulations (Levine, Toffel, and Johnson 2012). Yet even if a job does not directly damage a worker's body, it can take a mental toll, as in Frey's (1996) description of an air traffic controller who lost radio contact with the airplanes he was guiding:

> Watching in helpless horror as his planes careered farther and farther off course, the controller rose from his chair with an animal scream, burst into a sweat and began tearing off his shirt. By the time radio contact was re-established—and the errant planes were reined in—the controller was quivering on the floor half naked, and was discharged on a medical leave until he could regain his wits.

The costs of such workplace stress are potentially significant: The famous Whitehall studies (Marmot, Shipley, Hamilton 1978; Marmot et al. 1991) found an inverse relationship between employment rank in the British civil service and health outcomes, a pattern that has been interpreted as indicating that stress and other negative psychological features of low-ranking jobs may harm health, given that most of the study subjects at all ranks were office workers who had equal access to the National Health Service (Smith 1999).

Another nonpecuniary cost of many jobs that has attracted recent attention is inflexibility. Jobs in the so-called "gig economy" may provide more flexibility—for example, Uber drivers can decide exactly when they want to work, instead of taking shifts assigned by a manager (Hall and Krueger 2018), though potentially at the cost of reducing workers' wages or their ability to work full time when they wish to do so (Katz and Krueger 2016). Some recent studies have tried to quantify the value of flexibility by eliciting willingness to pay for increased autonomy in hours of work (Mas and Pallais 2017).

The main challenge we face in studying the aggregate changes in the nonpecuniary costs and benefits of work is that the survey data on workers' feelings are available only in 2010, 2012, and 2013. Thus, we must impute how workers felt about their jobs in past years based on recent information. Our strategy begins by measuring changes in the distribution of occupations. We then use the American Time Use Survey data to assign each occupation a vector of scores based on the feelings that its workers report in the recent data. Finally, we ask: If the distribution of occupations were different from what we see today, but feelings about each of the occupations stayed the same, how would workers' total experiences change? How much more or less stress, for example, would the workforce collectively experience if the distribution of occupations was the one observed in 1950, rather than the distribution observed today? How much more or less meaning, happiness, tiredness, and pain?

Our focus throughout is on market work. In the conclusion, we discuss how our approach might also be used to assess the consequences of women's significant increase in labor force participation in the postwar period, which would require a measure of feelings about nonmarket work.

Our approach relies on three key assumptions. First, we assume that the feelings an occupation produces today are the same as those it produced in the past. Second, although economists since Adam Smith (1776) have appreciated that pay may vary inversely with the nonpecuniary qualities of a job, we assume that any such compensating differentials do not affect the feelings that workers report. Third, we assume that the feelings a particular worker reports on the job are caused by that worker's occupation and not by his or her other circumstances or personality. At the end of the paper, we describe the implications of these assumptions for our findings as well as some robustness checks. The paper concludes by discussing some potential extensions and broader implications of our work, in particular how findings about the nonpecuniary characteristics of work should influence the analysis of big-picture labor market outcomes such as labor force participation and inequality.

Evolution of Occupations over Time

The first step in our estimation strategy is to measure shifts in the distribution of occupations over time. This requires us to categorize occupations in a consistent way in data from 1950 to the present day.

We use data from decennial Censuses from 1950 through 2000 and the 2011–2015 American Community Survey (ACS) to measure the distribution of occupations by sex, race, and education. Our occupation categories use the OCC1990 occupation coding produced by IPUMS (Integrated Public Use Microdata Series), described at https://usa.ipums.org/usa-action/variables/OCC1990. OCC1990 is based on the occupation codes used in the 1990 Census; it maps occupation codes used in other years to the 1990 codes, and aggregates some categories to make the coding more consistent over time. The American Time Use Survey data include only the current occupation coding scheme, which we map to OCC1990 ourselves.

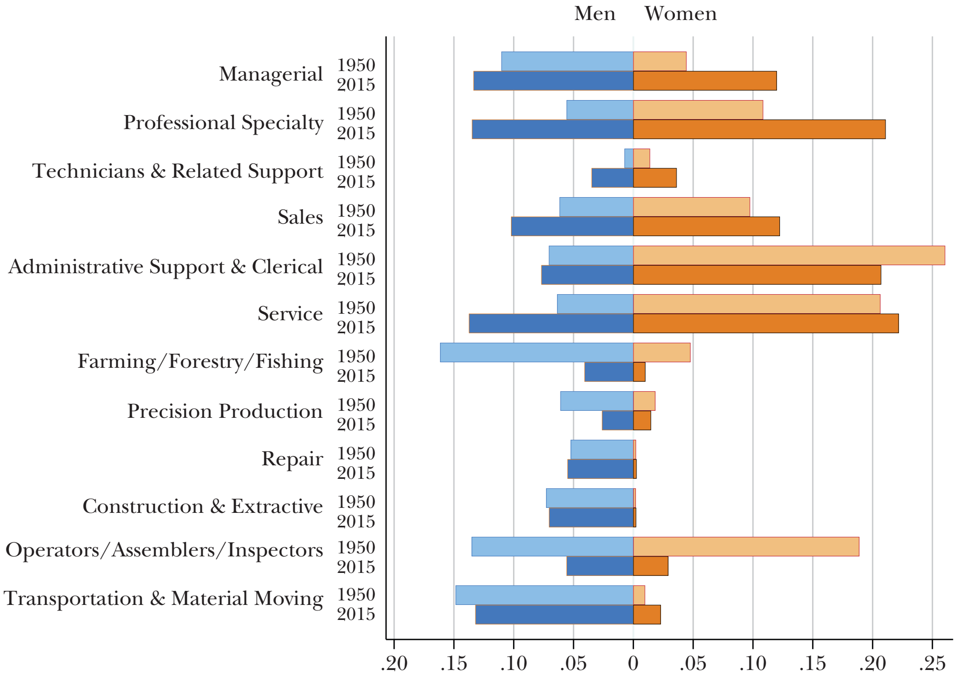
The OCC1990 coding contains 389 occupation categories. Some of these categories are so narrow that we observe very few workers in them in the American Time Use Survey—too few to be able to estimate feelings precisely for these occupations. In addition, even though OCC1990 is harmonized, it is not entirely uniform over time because of changes in the level of detail in the census occupation variables.[1] To improve the uniformity of the coding and to ensure a reasonably large number of people are used to calculate workers' feelings in each occupation, we aggregate the occupations to 12 broad categories. (We exclude military occupations.) Of course, aggregating occupations in this way poses the risk that the occupations categorized as, say, "sales occupations" in 1950 are quite different from those categorized as sales occupations in recent years. However, in analyses not reported here, we have found that we obtain similar overall results if we use the detailed OCC1990 codes, but the changes in the share of workers in each occupation become difficult to interpret (for example, because of the reclassification of detailed categories between 1950 and 1990).

With this coding in hand, we estimate the distribution of occupations by race, sex, and education in the 1 percent sample of the 1950 Census, the 5 percent samples of the 1960, 1980, 1990, and 2000 Censuses; the 1 percent form 1 and form 2 state samples of the 1970 Census; and the 2011–2015 five-year ACS sample. We obtain all datasets from IPUMS (Ruggles, Genadek, Goeken, Grover, and Sobek 2015). We also consider three education groups: a high school diploma or less, some college, and a bachelor's degree or more.

Figure 1 shows the categories and the distributions of men and women across occupations in 1950 and in 2015. Since 1950, both men and women have moved into managerial and professional specialty occupations, and out of farming and machine operating (the Operators/Assemblers/Inspectors category on the figure). Women have moved out of administrative support, but the share of men in that field has remained roughly constant. By contrast, men have shifted in large numbers into service occupations, while the share of women in service occupations is little

[1] For example, in the 1950 Census, almost all people in management jobs were recorded as "Managers, officials, and proprietors (not elsewhere classified)," which maps to the OCC1990 code "Managers and administrators, n.e.c." (code 022). But by the 1990 Census, which forms the basis for the OCC1990 codes, some managers were recorded as working in specialties, such as "managers of food-serving and lodging establishments" (code 017). Thus, a restaurant manager would be assigned the OCC1990 code 022 in the 1950 Census but code 017 in the 1990 Census or the 2011–2015 American Community Survey (ACS).

Figure 1
The Distribution of Workers across Broad Occupation Groupings in 1950 and 2015



Source: Authors' calculations from 1950 Census and 2011–2015 American Community Survey.
Note: The figure shows the proportion of men and women in each occupational grouping in 1950 and 2015.

changed. Some occupations, such as construction, have stable shares of the population over time.

These shifts create the potential for heterogeneity by sex in how feelings about work has changed, for two reasons. First, men and women have moved into and out of different occupations, so even if men and women don't differ from each other in their feelings about each occupation, the aggregate changes they have experienced will differ. Second, men and women may feel differently about the same occupations, so even where they have experienced similar changes in occupation shares, as with the shift into professional specialty occupations, the impact on the utility or disutility of work may differ. Our methodology will allow for both of these possible sources of change.

Feelings about Work

The American Time Use Survey, produced by the US Census Bureau, is a stratified random sample of the US population ages 16 and older. (Specifically, respondents to the survey are a subset of respondents to the Current Population

Survey.) The survey asks respondents to report, in significant detail, how they spent each minute of a day. Respondents also report their occupation in their main job (but not in any other jobs they may have).

In 2010, 2012, and 2013, the American Time Use Survey contained a "well-being module" that randomly selected three activities during the day for each respondent and asked the respondents to report their feelings while engaged in these activities. Activities were eligible to be randomly selected for these questions if they lasted at least five minutes and did not fall into the categories of sleeping, grooming, personal activities, refusal, or don't know. For the chosen activities, respondents were asked about how they were feeling during these activities along six dimensions: how happy, how sad, how stressed, how tired, how much pain, and how meaningful. They were asked to rank each of their feelings on a scale from 0 (not at all) to 6 (very much).[2]

To produce an index of feelings by occupation, we run six ordinary least squares regressions. In each case, the dependent variable is one of the six measures of feelings about time spent at work. The explanatory variables are 12 categories of occupations described in the previous section, and dummy variables for age, race, and education level. We then compute the mean response to each question within each occupation category, adjusted for differences in demographics across occupations. In online Appendix A1, we provide a detailed description of our procedure, including a table of the adjusted mean feelings for men and women in each occupation.

In our main analysis, we use these data only for respondents who were asked to report their feelings during the activity of working on their main job, but in online Appendix A2 we report results from an analysis with individual fixed effects that also uses data on feelings during activities other than work.

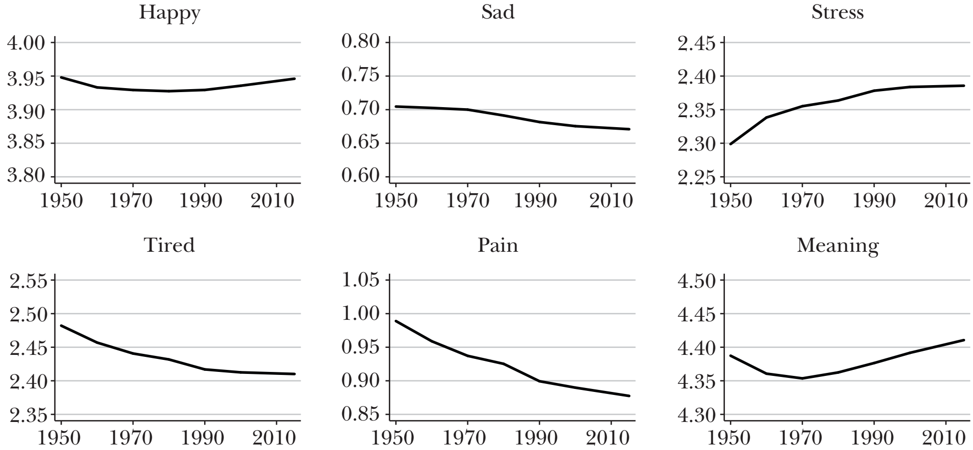
Our next step is to combine the (adjusted) estimates of feelings about work for each occupation with the census data on occupations described earlier. For each census year, we compute the average stress of work by taking a weighted average of the stress indexes of each occupation, weighting by the distribution of occupations in that year. We repeat this calculation for the five other feelings.[3]

Figure 2 shows how aggregate mean feelings at work have evolved over time for the six types of feelings in the data. Relative to 1950, the current distribution

[2] We obtain the American Time Use Survey (ATUS) microdata from the American Time Use Survey Data Extract Builder at http://www.atusdata.org (Hofferth, Flood, and Sobek 2015). We use the well-being module activity-level weights for estimation and normalize the weights such that the 2010, 2012, and 2013 samples receive equal weight in the calculations.

[3] In principle, to compute aggregate mean feelings in the present era, we could directly calculate means of self-reported feelings on the job in the American Time Use Survey (ATUS). However, because the ATUS is relatively small and not all respondents are asked to report their feelings at their main job, the distribution of occupations among ATUS respondents who report feelings on their main job could randomly differ by a significant amount from the population distribution of occupations. To rule out this problem, we estimate aggregate mean feelings in the present era with a weighted average of occupation-specific feelings, weighted by the occupation distribution in the 2011–2015 American Community Survey (ACS). See online Appendix A1 for details.

Figure 2
Changes in Aggregate Feelings at Work, 1950 to Present



Source: Authors' calculations from Census, American Community Survey, and American Time Use Survey.
Note: Lines show average of occupation scores weighted by distribution of occupations in each year. Occupation scores and occupation distributions are calculated for the full population; occupation scores are adjusted for age, race, sex, and years of education.

of occupations makes workers less sad and less tired, and makes them experience less pain. However, work also produces more stress. Happiness and meaningfulness both fell in early years, then rose in later years.[4]

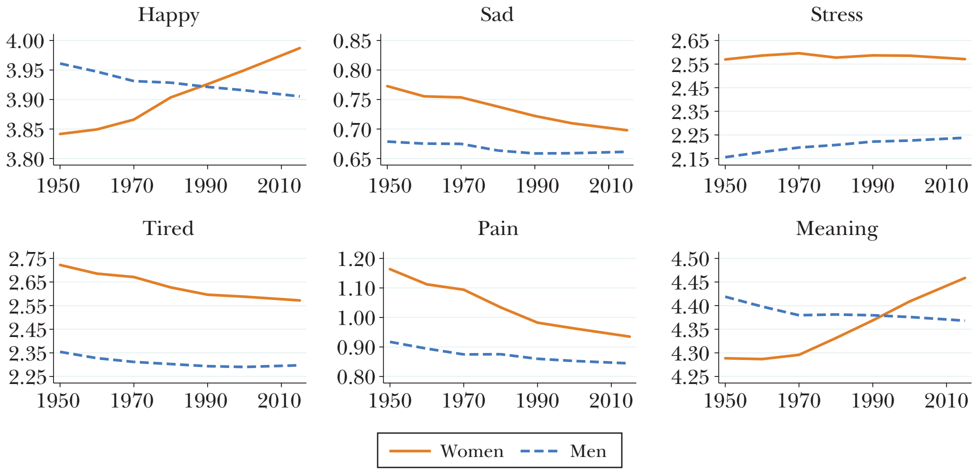
Feelings about Work by Gender, Education Level, and Race

The results change substantially when we calculate aggregate mean feelings separately by sex. To do this, we follow the same approach but compute both the adjusted mean feelings within each occupation category, and the yearly distribution of occupations, separately by demographic group (sex, sex × education, and sex × race × education). Online Appendix A1 provides details.

For women, the story is one of consistently improving feelings about work. Over time, as Figure 3 shows, work produced more happiness and a greater sense of meaning, and less sadness, tiredness, and pain; stress levels stayed roughly constant. Thus, over the period we examine, not only were women moving into the work

[4] It should be noted that estimation uncertainty in this calculation arises both from uncertainty in the estimation of the occupation shares and uncertainty in the estimation of the occupation-adjusted mean feelings. But in practice the census data are large enough that the occupation shares are estimated quite precisely, and uncertainty in the estimates of mean feelings by occupation in the American Time Use Survey data is the main source of uncertainty in our results.

Figure 3
Changes in Aggregate Feelings at Work by Sex, 1950 to present



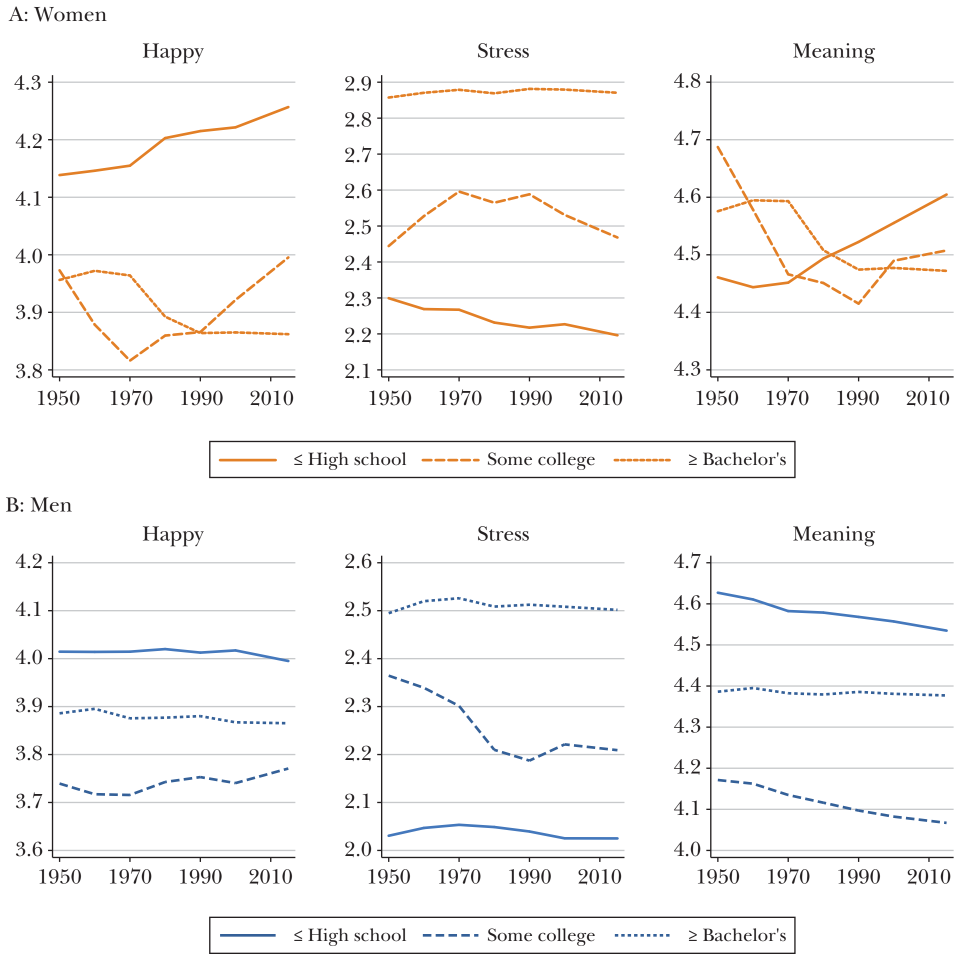
Source: Authors' calculations from the Census, American Community Survey, and American Time Use Survey.
Note: Lines show average of occupation scores weighted by distribution of occupations in each year. Occupation scores and occupation distributions are calculated separately by sex; occupation scores are adjusted for age, race, and years of education.

force but they also were shifting to occupations with better nonpecuniary attributes. By contrast, for men, the picture is more mixed: Although work became less painful and tiring, it also became more stressful, less meaningful, and less happy.

These patterns may be partly the result of aggregating together very different occupations, such as including both restaurant managers and chief executive officers in the managerial category. We can attempt to classify occupations more finely, despite the lack of perfectly uniform coding across years, if we divide the sample by education. When we do this, we re-estimate the adjusted mean feelings within each occupation, using only data on workers at a given education level (as explained in online Appendix A1).

Figure 4 shows how happiness, stress, and meaningfulness have evolved when we divide the sample by both education and sex. We concentrate on happiness, stress, and meaning because there appears to be little interesting heterogeneity in tiredness and pain, and sadness appears to be the inverse of happiness. For women, shown in the top panel of the figure, the gains in happiness and meaningfulness are concentrated among those with no more than a high school diploma. The highest-educated women actually show falling happiness and meaningfulness, similar to the overall findings for men. For men, shown in the bottom panel, there is a clear drop in meaningfulness at lower education levels. However, we find no rise in stress and little decrease in happiness for men within education groups, suggesting that the trends in these variables for men overall might result from men at different

Figure 4
Changes in Aggregate Feelings at Work by Sex and Education



Source: Authors' calculations from the Census, American Community Survey, and American Time Use Survey.
Note: Lines show average of occupation scores weighted by distribution of occupations in each year. Occupation scores and occupation distributions are calculated separately by sex and education; occupation scores are adjusted for age, race, and years of education.

education levels having different feelings about the same occupation, which is not detected when we look at men overall.

These conclusions should be regarded as tentative. One reason is that disaggregating by education means that we are using a smaller sample to estimate each occupation score. Another reason is that disaggregating by education could under some circumstances exacerbate rather than reduce any bias in our estimates of feelings by occupation—rising education levels within occupations (for example, see

Spitz-Oener 2006) mean that the type of managerial work done by someone with a high school diploma today may be quite different from that done by someone with a high school diploma in 1950.

We can also disaggregate the results by race. We examine only whites and blacks because the sample size of the American Time Use Survey contains too few respondents of other races to obtain precise estimates when we disaggregate by race, sex, and education. In Figure 5, we focus on estimates for people with a high school education or less. The trends in meaningfulness are the same across races—meaningfulness has risen for both white and black women, and fallen for both white and black men. However, happiness has risen for white women while falling for black women, and stress has risen for black men while falling for white men and women. Importantly, these estimates account only for differences in feelings about the occupation itself, not for differences in pay. If racial discrimination in pay varies across occupations or has changed over time, the change in workers' overall happiness could be quite different.

Sources of the Shifts

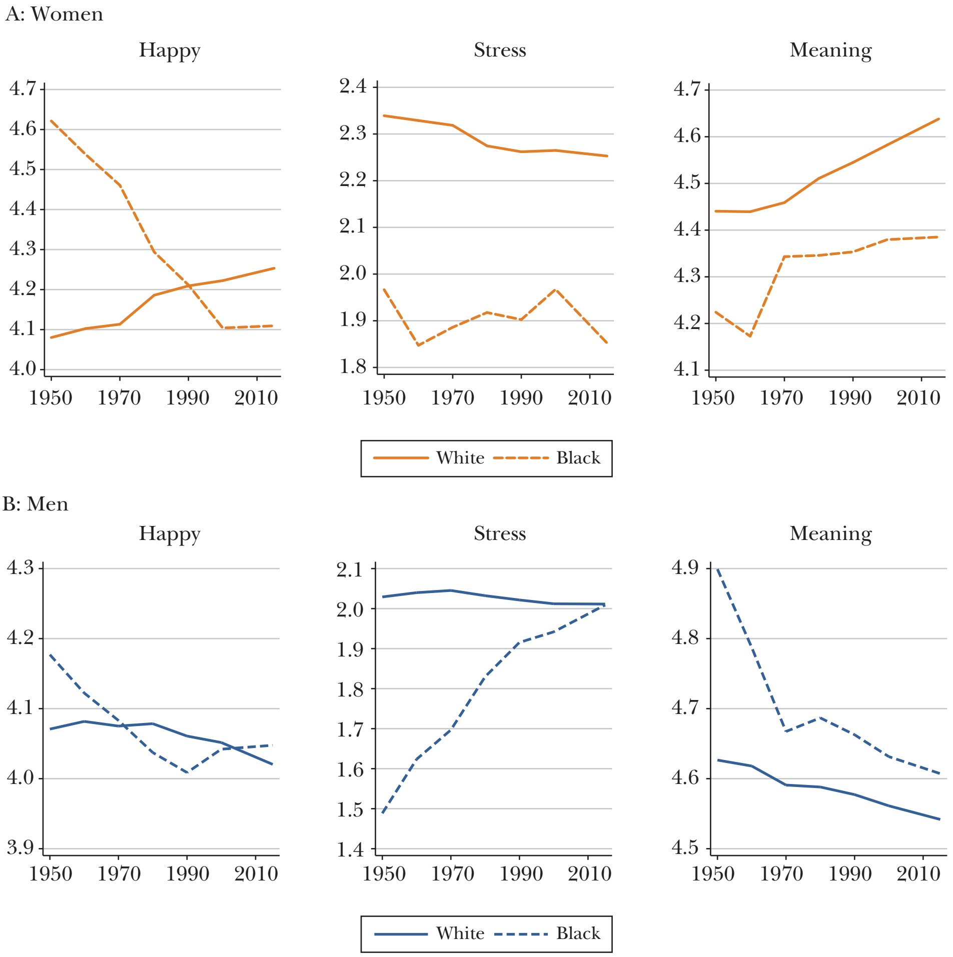
What is driving these shifts in aggregate feelings about work? To gain some insight into this issue, we plot the relationship between an occupation's average feelings and the change in the share of workers in that occupation since 1950. Figure 6 shows these relationships for happiness, stress, and meaning, separately for women and men. Each circle in the graphs represents a different occupation category, with area proportional to the occupation's share of workers in 1950.

The figure shows that the different results for men and women arise not only from differences in how their occupation distributions have changed, but also from differences in the feelings they report in the same occupational categories. For example, both men and women are less likely now than in 1950 to work as machine operators, assemblers, and inspectors. For women, such jobs are associated with below-average happiness and meaningfulness, so the shift increases women's happiness and meaning at work. For men, such jobs are associated with *above*-average happiness and meaningfulness, as well as below-average stress, so the same shift in the occupation distribution *decreases* the nonpecuniary value of work for men.

These patterns suggest that the overall improvements for women appear to be driven by their shift into professional and managerial work and out of factory work, while the overall decreases for men appear to be driven by their shift out of farming and factory work and into professional and service occupations.

We emphasize that the gender differences in reported feelings for the same occupation can be interpreted in multiple ways. The differences could mean that men and women feel differently about exactly the same jobs. Alternatively, within a single occupation code, men and women might on average be doing slightly different jobs that our coding is not sufficiently detailed to reveal. Another possibility is that men and women feel the same about the actual tasks involved in the work but that the

Figure 5
Changes in Aggregate Feelings at Work by Sex and Race (Education ≤ High School), 1950 to the Present



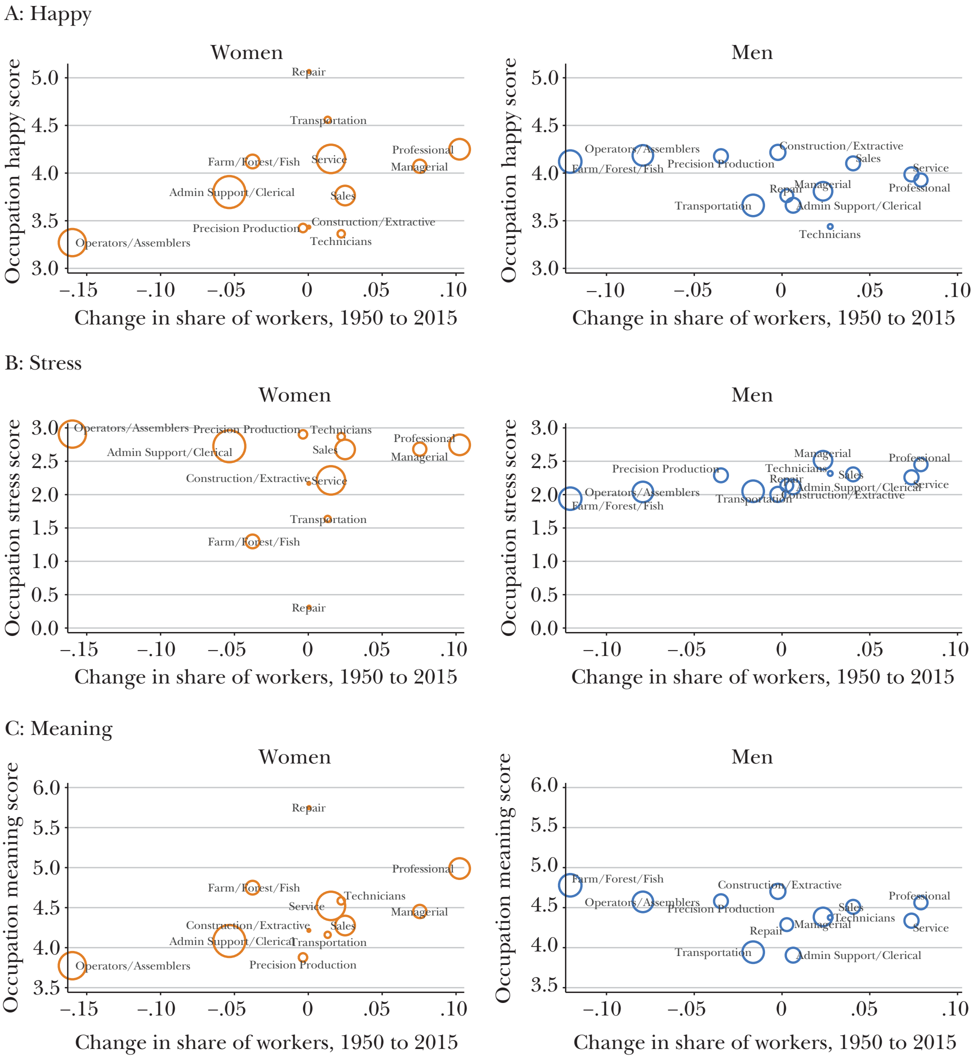
Source: Authors' calculations from the Census, American Community Survey, and American Time Use Survey.
Note: Lines show average of occupation scores weighted by distribution of occupations in each year. Occupation scores and occupation distributions are calculated for black and white respondents with no more than a high school education, separately by race and sex; occupation scores are adjusted for age and years of education.

broader work environment has disparate impacts on men and women, such as when sexual harassment occurs. Finally, to the extent that feelings about work are socially constructed, the reported gender differences might reflect messages that society sends to men and women about how they "should" feel about different jobs, rather than any differences in how people would feel absent such messages.

Figure 6

Changes in Occupation Share and Average Feelings by Occupation

(area of circles proportional to share of workers in occupation in 1950)



A: Happy

B: Stress

C: Meaning

Source: Author's calculations from the Census, American Community Survey, and American Time Use Survey.

Note: Occupational scores and occupational distributions calculated separately by sex; occupation scores adjusted for age, race, and education.

Discussion of Assumptions

We relied on several strong assumptions to measure workers' feelings about different occupations. Here, we discuss the potential biases that these assumptions may create and some robustness checks that we have carried out to help evaluate their importance.

Have Feelings about Occupations Changed over Time?

Our approach assumes that the feelings an occupation produced in the past are the same as those it produces today. Direct measures of workers' feelings about their job in past eras would be preferable if they were available, but we are not aware of any historical data on subjective feelings about work in particular occupations that we can measure in a consistent way over time.

This assumption may be particularly problematic when considering the physical costs of work. Regulations and technological improvements have made many jobs safer than in the past. In addition, in many occupations, capital equipment has substituted for human effort; a miner today is more likely to operate heavy machinery and less likely to wield a pickaxe.

To the extent that changes in occupations over time have merely dampened the differences between occupations but not erased or reordered them, our results will underestimate the effect of changes in the occupation distribution. For example, if working on an assembly line has always been more tiring than working as retail sales clerk, but the gap is smaller today than in the past, our approach will underestimate the difference in tiredness between manufacturing and retail jobs in 1950 and therefore underestimate the change in average feelings caused by the large-scale shift from the former occupation to the latter.

However, it is also possible that the relative rank of occupations has changed over time. For example, the "meaningfulness" of a particular occupation may in part be socially constructed and depend on the value that the worker's family, friends, neighbors, or society at large happen to place on that occupation. Workers in various occupations have sought to increase their prestige by defining them as professions (Larson 1977), and public opinion polls have also measured fluctuations in occupations' prestige over time (for example, Taylor 2001). Absent data on workers' feelings in the past, we cannot assess how such changes might have affected our results.

A related possibility is that the relationship between education and how workers feel about their jobs has varied across cohorts. For example, the education system might somehow encourage people who reach a certain schooling level to view a particular kind of occupation as especially appropriate or meaningful, but which occupations these are might change over time. In online Appendix A3, we investigate this possibility by controlling for an interaction of age and education when estimating the average feelings in each occupation. This change has little impact on our results.

In addition, it is possible that workers' preferences for particular attributes of work have changed over time, for example, perhaps because preferences have been altered as a result of rising incomes since 1950. Again, absent data on past feelings, we have no way to measure the direction or magnitude of changes in preferences, although one might extend our approach by seeking to estimate the relationship between income and cross-occupation differences in feelings.

What about Compensating Differentials?

We do not attempt to measure differences in wages between occupations or how these differences might relate to the feelings that workers report. This focus means that, at most, our analysis provides an account of how the changing occupation distribution affects nonpecuniary costs and benefits of work, but cannot hope to describe the effect on workers' overall welfare.

Moreover, our calculation assumes that the feelings workers report in the American Time Use Survey depend only on the actual jobs they are doing and not on how much they are paid to do those jobs, because our method takes account of changes over time in the distribution of occupations but not of changes in the wage for each occupation. If, instead, the feelings that workers report in the survey also depend on their pay, the compensating differentials would likely lead us to underestimate the differences between occupations and underestimate the effect of changes in the occupation structure. For example, if workers who are less happy about their tasks are also more happy about their wages, then the reported difference in happiness between two occupations would be biased downward. This bias is likely of greatest concern for dimensions such as happiness and sadness that potentially reflect workers' overall views of a job, and is of less concern for dimensions that measure more specific feelings such as meaning or pain.

Related, our results treat each dimension of feelings about work as separate and do not attempt to map changes in the vector of feelings into changes in a single index of the amount of (dis-)utility that workers experience. One could compute such an index by estimating compensating differentials for the feelings that different occupations produce—via hedonic wage models of the type pioneered by Tinbergen (1956)—and then calculating the compensating variation associated with a change in the occupation distribution. However, estimating the compensating differentials is not straightforward, due to the way workers with heterogeneous preferences endogenously sort across occupations (Bartik 1987; Epple 1987). We leave such calculations for future research.

Might Feelings about Occupations Reflect other Individual Traits?

Our approach assumes that the feelings a particular worker reports on the job are caused by that worker's occupation, rather than by other circumstances or personality. For example, when we observe that people in managerial occupations report an above-average level of stress, we assume that this is because management work is inherently stressful, and not because people who would feel stressed in any job are more likely to end up being managers. This assumption will fail if occupation choices are correlated with other factors that affect a person's feelings, and if it fails, our results will be biased.

One possible robustness check to address this issue, using the American Time Use Survey data, is to control for the feelings that workers report when they are not on the job. These feelings in nonwork activities might be viewed as an indicator of the permanent feelings that a person would report regardless of occupation, so controlling for them might adjust for all of the nonoccupation differences between respondents. In online Appendix A2, we carry out this calculation, using a fixed-effects estimator to identify the effect of an occupation by measuring the difference between the feelings that a worker reports on the job and the feelings that the very same worker reports in other activities. We refer to our original measure of the feelings generated by an occupation as the occupation's baseline score, and to the measure that is adjusted for feelings reported during other activities as the occupation's fixed-effects score.

For most occupations whose shares changed substantially, and for most of the types of feelings that we measure, the fixed-effects and baseline occupation scores are closely correlated. This correlation gives some confidence that our basic approach to measuring the feelings induced by an occupation is reasonable.

However, there are a few outliers in the occupation scores, which can lead to different estimates of the aggregate trend in feelings. For example, among women, the fixed-effects score for machine operators, assemblers, and inspectors often differs substantially from the baseline score. This occupation was one of the lowest scoring on happiness for women in the baseline but one of the highest scoring on happiness for women in the fixed effects estimate. Also, this occupation shrunk substantially from 1950 to 2015. As a result, women in this occupation had a downward trend in happiness according to the fixed effects estimates, but an upward trend according to the baseline estimates. For men, farming, forestry, and fishing were rated very high in meaning in the baseline estimates but quite low in the fixed effects estimates, while service occupations received a moderate meaning score in the baseline and a high score with fixed effects. Thus, the shrinkage of the agricultural sector and the growth of service work implied decreasing meaningfulness for men according to the baseline estimates but rising meaningfulness according to the fixed effects estimates. The differences between the fixed effects and baseline estimates appear to be concentrated at lower education levels.

The differences between the fixed effects and baseline estimates suggest a need for caution in interpreting the overall results. However, the fixed-effects approach is imperfect in various ways and might have biases of its own. For example, if a good job also gives the worker more positive feelings when she is at home, the difference in feelings between work and home will underestimate the true effect of the occupation on how the worker feels. Alternatively, if people who have bad jobs are particularly happy to go home from them, the difference in feelings between work and home will overestimate the effect of the occupation on feelings. As a result, the fixed-effects measures could be biased in either direction relative to the true change in feelings about work.

Conclusion

The distribution of occupations has changed significantly in the post–World War II period. People feel differently about different occupations, and in addition, people in different demographic groups appear to feel differently about the same occupations. Taking all of these factors together, we find substantial shifts both in the aggregate utility and disutility derived from work and in the distribution of that (dis-)utility across people.

Our work highlights how more measurement of the characteristics of work beyond income could offer insight on a number of large-scale questions. As one example, there has been a well-documented secular rise in wage and income inequality (Heathcote, Perri, and Violante 2010), in particular at the top of the distribution (Piketty and Saez 2003). Are rising wages at the top of the distribution (at least in part) a compensating differential for particularly demanding jobs, so that inequality in wages exceeds inequality in the total rewards of work? Or are the nonpecuniary benefits of work also increasingly concentrated at the top?

In addition, one of the biggest changes in the labor market in the postwar period has been the rise in women's participation. Yet little is known about the overall welfare consequences of this change because almost all research has assumed that wages are the sole benefit, and opportunity costs of time are the sole costs, of working. But as we show, men and women have different likelihoods of working in some occupations and sometimes feel differently on average about those occupations. Thus, we cannot simply extrapolate from the experience of men to calculate the costs and benefits of work for women. Furthermore, nonmarket work represents a large fraction of economic activity and of how people, especially women, spend their time (Aguiar and Hurst 2016; Waring 1988). Our approach could in principle be applied to measure feelings in nonmarket work and then analyze how aggregate feelings about *all* work—market and nonmarket—changed with shifts in labor force participation. We did not pursue that question because it is challenging to define a sharp boundary between nonmarket work and leisure in the data available to us, but such a study might help provide a more complete account of the implications of rising female labor force participation. Such an analysis could also consider changes in the utility of leisure time, following the study by Aguiar, Bils, Charles, and Hurst (2017) of how improvements in video games reduced employment among low-skilled young men.

Finally, our work provides a new perspective on an old question about the labor market: Why do people keep working full-time hours even as income levels have risen? (Or, for those concerned about whether robotics and artificial intelligence will lead to a sharp reduction in demand for labor, how much do people actually want to work?) In a famous essay, "Economic Possibilities for our Grandchildren," Keynes (1930) raised the possibility that as incomes rose, people would spend few hours working:

> For many ages to come … everybody will need to do some work if he is to be contented. We shall do more things for ourselves than is usual with the rich today, only too glad to have small duties and tasks and routines. But beyond this, we shall endeavour to … make what work there is still to be done to be as widely shared as possible. Three-hour shifts or a fifteen-hour week may put off

> the problem for a great while. For three hours a day is quite enough to satisfy the old Adam in most of us!

It is by now well known that, at least in a strict sense, this prediction of fewer hours worked has not yet been borne out. But Keynes' argument also hints at a possible reason why: Work is not motivated by wages alone. The prediction of fewer hours really should apply only to the component of work that produces disutility, not the component that produces positive utility. It is possible that, in the aggregate, people in wealthy countries do much less "work," in the sense of an activity that is a source of disutility, than in Keynes' time because more of the time spent working is associated with experiences that workers value positively. And it is also possible that these changes in the experience of working have been disproportionately felt in different parts of the income distribution and different demographic groups. However, there have not been attempts to measure changes in the aggregate nonwage aspects of work over time. This paper is a first small step in that direction.

■ *We thank Enrico Moretti, Gordon Hanson, and Timothy Taylor for very helpful comments and Sharada Dharmasankar for excellent research assistance. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System.*

References

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics*, vol. 4B, edited by Orley Ashenfelter and David Card. Amsterdam: Elsevier.

Aguiar, Mark, Mark Bils, Kerwin Kofi Charles, and Erik Hurst. 2017. "Leisure Luxuries and the Labor Supply of Young Men." NBER Working Paper 23552.

Aguiar, Mark, and Erik Hurst. 2016. "The Macroeconomics of Time Allocation." Chap. 5 in *Handbook of Macroeconomics*, vol. 2A, edited by John B. Taylor and Harald Uhlig. Amsterdam: Elsevier.

Aldrich, Mark. 1997. *Safety First: Technology, Labor and Business in the Building of Work Safety, 1870–1939.* Baltimore: Johns Hopkins University Press.

Ariely, Dan, Emir Kamenica, and Dražen Prelec. 2008. "Man's Search for Meaning: The Case of Legos." *Journal of Economic Behavior & Organization* 67(3–4): 671–77.

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

Bartik, Timothy J. 1987. "The Estimation of Demand Parameters in Hedonic Price Models." *Journal of Political Economy* 95(1): 81–88.

Case, Anne, and Angus Deaton. 2017. "Mortality and Morbidity in the 21st Century." *Brookings Papers on Economic Activity*, Spring, pp. 397–443.

Centers for Disease Control and Prevention. 1999. "Achievements in Public Health, 1900–1999: Improvements in Workplace Safety—United States, 1900–1999." *Morbidity and Mortality Weekly Report* 48(22): 461–69.

Epple, Dennis. 1987. "Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products." *Journal of Political Economy* 95(1): 59–80.

Frey, Darcy. 1996. "Something's Got to Give." *New Times Magazine*, March 24.

Grant, Adam M. 2007. "Relational Job Design and the Motivation to Make a Prosocial Difference." *Academy of Management Review* 32(2): 393–417.

Grant, Adam M. 2012. "Leading with Meaning: Beneficiary Contact, Prosocial Impact, and the Performance Effects of Transformational Leadership." *Academy of Management Journal* 55(2): 458–76.

Hall, Jonathan V., and Alan B. Krueger. 2018. "An Analysis of the Labor Market for Uber's Driver-Partners in the United States." *ILR Review* 71(3): 705–32.

Hamermesh, Daniel S. 2001. "The Changing Distribution of Job Satisfaction." *Journal of Human Resources* 36(1): 1–30.

Heathcote, Jonathan, Fabrizio Perri, and Giovanni L. Violante. 2010. "Unequal We Stand: An Empirical Analysis of Economic Inequality in the United States, 1967–2006." *Review of Economic Dynamics* 13(1): 15–51.

Herrendorf, Berthold, Richard Rogerson, and Akos Valentinyi. 2014. "Growth and Structural Transformation." Chap. 6 In *Handbook of Economic Growth*, vol. 2B, edited by Philippe Aghion and Steven N. Durlauf. Amsterdam: Elsevier.

Hofferth, Sandra L., Sarah M. Flood, and Matthew Sobek. 2015. American Time Use Survey Data Extract Builder: Version 2.5 [dataset]. University of Maryland and and University of Minnesota. http://doi.org/10.18128/D060.V2.5.

Kalleberg, Arne L. 2011. *Good Jobs, Bad Jobs: The Rise of Polarized and Precarious Employment Systems in the United States, 1970s to 2000s.* New York: Russell Sage Foundation.

Katz, Lawrence F., and Alan B. Krueger. 2016. "The Rise and Nature of Alternative Work Arrangements in the United States, 1995–2015." NBER Working Paper 22667.

Keynes, John Maynard. 1930. "Economic Possibilities for Our Grandchildren." Reprinted in *Essays in Persuasion* (1931), pp. 358–73. London: Macmillian and Co.

Larson, Magali Sarfatti. 1977. *The Rise of Professionalism: A Sociological Analysis.* Berkeley: University of California Press.

Levine, David I., Michael W. Toffel, and Matthew S. Johnson. 2012. "Randomized Government Safety Inspections Reduce Worker Injuries with No Detectable Job Loss." *Science* 336(6083): 907–11.

Marmot, M. G., G. Rose, M. Shipley, and P. J. S. Hamilton. 1978. "Employment Grade and Coronary Heart Disease in British Civil Servants." *Journal of Epidemiology and Community Health* 32(4): 244–49.

Marmot, M. G., S. Stansfeld, C. Patel, F. North, J. Head, I. White, E. Brunner, A. Feeney, M. G. Marmot, and G. Davey Smith. 1991. "Health Inequalities among British Civil Servants: The Whitehall II Study." *Lancet* 337(8754): 1387–93.

Mas, Alexandre, and Amanda Pallais. 2017. "Valuing Alternative Work Arrangements." *American Economic Review* 107(12): 3722–59.

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

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2015. Integrated Public Use Microdata Series: Version 6.0 [dataset]. University of Minnesota. http://doi.org/10.18128/D010.V6.0.

Schwartz, Barry. 2015. *Why We Work.* New York: Simon & Schuster.

Smith, Adam. 1776 [2003]. *The Wealth of Nations.* New York: Bantam Dell.

Smith, James P. 1999. "Healthy Bodies and Thick Wallets: The Dual Relation Between Health and Economic Status." *Journal of Economic Perspectives* 13(2) 145–66.

Spitz-Oener, Alexandra. 2006. "Technical Change, Job Tasks, and Rising Educational Demands: Looking Outside the Wage Structure." *Journal of Labor Economics* 24(2): 235–70.

Taylor, Humphrey. 2001. "Doctors the Most Prestigious of Seventeen Professions and Occupations, Followed by Teachers (# 2), Scientists (#3), Clergy (#4) and Military Officers (#5)." *The Harris Poll* no. 50. https://theharrispoll.com/wp-content/uploads/2017/12/Harris-Interactive-Poll-Research-DOCTORS-THE-MOST-PRESTIGIOUS-OF-SEVENTEEN-PROFESSI-2001-10.pdf.

Tinbergen, Jan. 1956. "On the Theory of Income Distribution." *Weltwirtschaftliches Archiv* 77: 155–75.

Waring, Marilyn. 1988. "If Women Counted: A New Feminist Economics." San Francisco: Harper & Row.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 259–280

Social Connectedness: Measurement, Determinants, and Effects

Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong

Social networks can shape many aspects of social and economic activity: migration and trade, job-seeking, innovation, consumer preferences and sentiment, public health, social mobility, and more. In turn, social networks themselves are associated with geographic proximity, historical ties, political boundaries, and other factors. Traditionally, the unavailability of large-scale and representative data on social connectedness between individuals or geographic regions has posed a challenge for empirical research on social networks. More recently, a body of such research has begun to emerge using data on social connectedness from online social networking services such as Facebook, LinkedIn, and Twitter. To date, most of these research projects have been built on anonymized administrative microdata from Facebook, typically by working with coauthor teams that include Facebook employees. However, there is an inherent limit to the number of researchers that will be able to work with social network data through such collaborations.

In this paper, we therefore introduce a new measure of social connectedness at the US county level. Our Social Connectedness Index is based on friendship links on Facebook, the global online social networking service. Specifically, the Social Connectedness Index corresponds to the relative frequency of Facebook friendship

■ *Michael Bailey is a Senior Economic Research Scientist, Facebook, Inc., Menlo Park, California. Rachel Cao is a PhD candidate in economics, Harvard University, Cambridge, Massachusetts. Theresa Kuchler is an Assistant Professor of Finance and Johannes Stroebel is a Professor of Finance, both at the Stern School of Business, New York University, New York City, New York. Arlene Wong is an Assistant Professor of Economics at Princeton University, Princeton, New Jersey. Stroebel is the corresponding author at johannes.stroebel@nyu.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.259 doi=10.1257/jep.32.3.259

links between every county-pair in the United States, and between every US county and every foreign country. Given Facebook's scale, with 2.1 billion active users globally and 239 million active users in the United States and Canada (Facebook 2017), as well as the relative representativeness of Facebook's user body, these data provide the first comprehensive measure of friendship networks at a national level. Moreover, the Social Connectedness Index data can be made accessible to members of the broader research community. Interested researchers are invited to email sci_data@fb.com to learn about the current process for working with the Social Connectedness Index data.

We begin this article by describing the construction of the Social Connectedness Index (SCI). The bulk of the paper then explores various patterns related to social connectedness. We first use the SCI data to analyze patterns of social connectedness between US counties. We find that the intensity of friendship links is strongly declining in geographic distance, with the elasticity of the number of friendship links to geographic distance ranging from about –2.0 over distances less than 200 miles, to about –1.2 for distances larger than 200 miles. We also look at how social connectedness is shaped by political boundaries such as state lines, exposure to large within-US population movements, and other historical and contemporaneous factors.

We then explore heterogeneity across counties in the geographic concentration of their populations' social networks. For the average county, 62.8 percent of all friendship links are to individuals living within 100 miles, but this number ranges from 46.0 percent at the 5th percentile to 76.9 percent at the 95th percentile of the across-county distribution. We find that the populations of counties with a larger fraction of friends living more than 100 miles away are on average better off along a number of socioeconomic dimensions. For example, counties with more geographically dispersed social networks have higher incomes, higher education levels, and higher social mobility.

We then turn to the question of how the intensity of social connectedness between regions correlates with bilateral economic and social activity. We first document a strong correlation between social connectedness and trading activity, consistent with recent research that argues that social networks help overcome informational and cultural frictions that can inhibit trade. Social connectedness is also positively correlated with the spread of innovation and within-US migration. When we look at friendship links between US regions and foreign countries, we find further strong correlations with both past migration patterns and present-day trade flows.

Throughout this essay, our focus is on documenting and describing salient patterns of social connectedness across a variety of settings. We do not seek to provide causal analyses, nor do we want to imply causal relationships behind the correlations we document. Nevertheless, we do believe that our findings can guide future research on the causal effects of social networks. More generally, the patterns discussed here highlight significant opportunities for using data from online social networking services such as Facebook to help alleviate the measurement challenges faced by researchers across the social sciences trying to better understand the role of social connectedness.

Measuring Social Connectedness

The Social Connectedness Index is constructed using aggregated and anonymized information from the universe of friendship links between all Facebook users as of April 2016. Duggan, Ellison, Lampe, Lenhart, and Madden (2015) report that as of September 2014, more than 58 percent of the US adult population and 71 percent of the US online population used Facebook. The same source reports that, among online US adults, Facebook usage rates are relatively constant across income groups, education groups, and racial groups. Usage rates among online US adults are declining in age, from 87 percent of 18-to-29 year-olds to 56 percent of above-65 year-olds.

In the United States, Facebook mainly serves as a platform for real-world friends and acquaintances to interact online, and people usually only add connections on Facebook to individuals whom they know in the real world (Jones et al. 2013; Gilbert and Karahalios 2009; Hampton, Goulet, Rainie, and Purcell 2011). Establishing a friendship link on Facebook requires the consent of both individuals, and the total number of friends for a person is limited to 5,000. As a result, Facebook data have a unique ability to provide a large-scale representation of US friendship networks.

To measure the social connectedness between geographies, we map Facebook users to their respective county and country locations, and thus obtain the total number of friendship links between these geographies. Locations are assigned to users based on the users' information and activity on Facebook, including the stated city on their Facebook profile, and device and connection information. We only consider friendship links among Facebook users who have interacted with Facebook over the 30 days prior to the April 2016 snapshot.[1] We treat each friendship link identically.

We then construct the Social Connectedness Index between all pairs of 3,136 US counties, and between every US county and every foreign country, as the normalized total number of friendship links for each geographic pair. In particular, the Social Connectedness Index is constructed to have a maximum value of 1,000,000, and relative differences in the index correspond to relative differences in the total number of friendship links. The highest Social Connectedness Index value of 1,000,000 is assigned to Los Angeles County–Los Angeles County connections (Los Angeles County is where people have the most friends with other people in their county).

The Determinants of Social Connectedness

The Social Connectedness Index can be used to analyze the correlates of the intensity of social connectedness between US counties. We first analyze the role

[1] Facebook formally defines such "monthly active users" in its 10Q statements as follows: "We define a monthly active user as a registered Facebook user who logged in and visited Facebook through our website or a mobile device, or used our Messenger application (and is also a registered Facebook user), in the last 30 days as of the date of measurement."

of geographic distance in shaping social connectedness in the United States. The effects of geographic proximity on friendship formation and social interactions have been studied in a number of papers, including Zipf (1949), Verbrugge (1983), and Marmaros and Sacerdote (2006).

As a motivating example, compare San Francisco County and Kern County in California. These two counties have roughly the same population of slightly under one million, but Kern County is 175 times larger in area. Moreover, San Francisco County, which is home to the city of San Francisco, is surrounded by the urbanized Bay Area economy including Oakland and San Jose. Kern County includes the Bakersfield metro area, but it is not surrounded by an urban area.

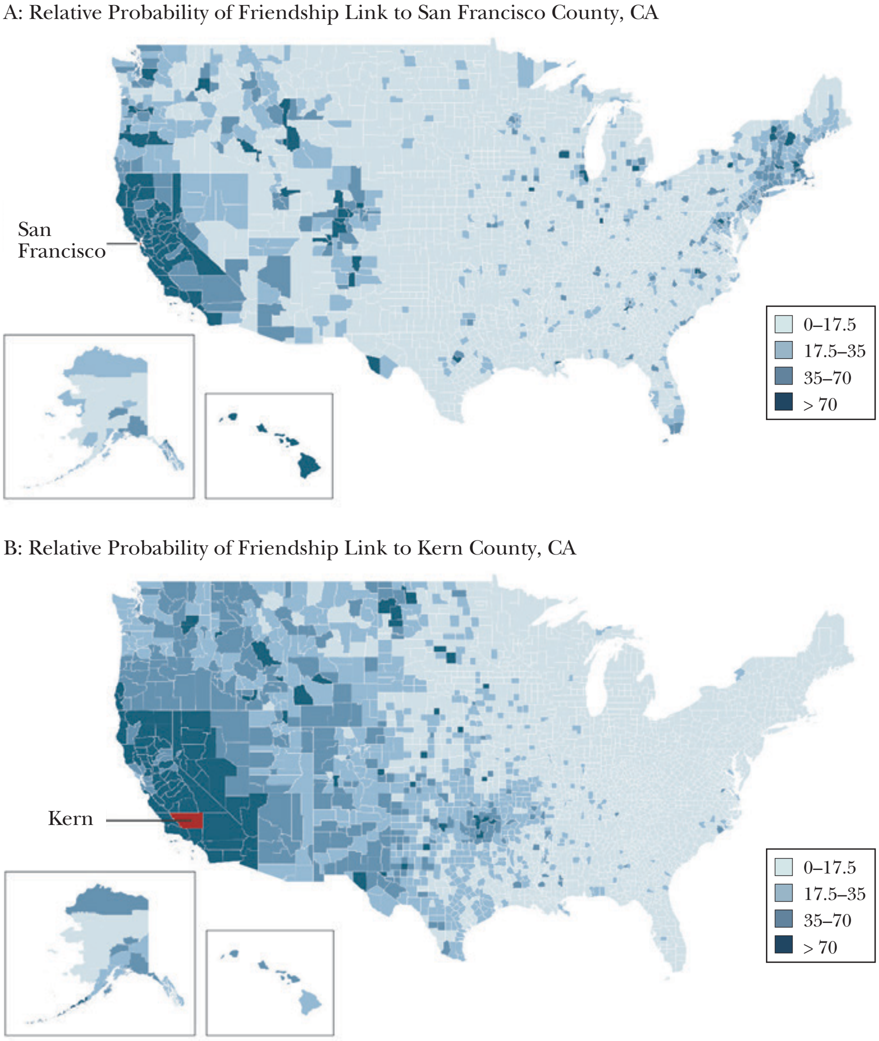
We construct a measure that we call the "relative probability of friendship" by taking the Social Connectedness Index between counties *i* and *j* and dividing it by the product of the number of Facebook users in the two counties. This allows us to take into account the fact that we will see more friendship links between counties with more Facebook users.[2] If this measure is twice as large, this means that a given Facebook user in county *i* is about twice as likely to be connected with a given Facebook user in county *j*. The heat maps in Figure 1 show the relative probability that a given Facebook user in San Francisco County (Figure 1A) or Kern County (Figure 1B) is connected to a given Facebook user in another county.

For both San Francisco County and Kern County, a significant proportion of friendship links (dark shading indicates more links) are to geographically close counties across the West Coast. However, there are also noticeable differences in the social connectedness of the two counties. The population of San Francisco County has significant social connections to counties located in the northeastern United States, while the population of Kern County has far fewer of these friendship links. Instead, Kern County's friendship network is very concentrated in the West Coast and Mountain States, with the exception of a pocket of strong connections to individuals living in Oklahoma and Arkansas. These connections are likely related to past migration patterns, because Kern County was a major destination for migrants fleeing the Dust Bowl in the 1930s. Kern County also has substantial friendship links to the oil-producing regions of North Dakota, perhaps not surprising given that Kern County produces more oil than any other county in the United States.

Overall, the friendship networks of the Kern County population are much more geographically concentrated than those of the San Francisco County population: Kern County has 57 percent of friends living within 50 miles, relative to 27 percent for San Francisco County. In comparison with the summary statistics for the whole United States, displayed in Table 1, the geographic concentration of the friendship network of Kern County is similar to the US average while San Francisco County's friendship network is extremely geographically dispersed. For the average (population-weighted) US county, 55.4 percent of friends live within 50 miles, with a

[2] While the number of Facebook users per county is not part of the public data release, very similar patterns for "relative probability of friendship" would be obtained if we instead divided the Social Connectedness Index by the product of county-level populations.

Figure 1
County-Level Friendship Maps


A: Relative Probability of Friendship Link to San Francisco County, CA

B: Relative Probability of Friendship Link to Kern County, CA

Note: The heat maps show the relative probability that a Facebook user in each county *j* has a friendship link to San Francisco County, CA (Panel A) and Kern County, CA (Panel B). Darker colors correspond to counties in which there is a higher probability of a friendship link between a person in home county *i* (San Francisco or Kern) and county *j*. The "relative probability of friendship" is constructed by taking the Social Connectedness Index between counties *i* and *j* and dividing it by the product of the number of Facebook users in the two counties.

10–90 percentile range of 42.5 to 67.4 percent; and over 70 percent of friends live within 200 miles, with a 10–90 percentile range of 57.1 to 81.2 percent. This despite the fact that, for the average county, only 1.3 percent and 6.6 percent of the US population live within 50 miles and 200 miles, respectively.

Table 1
Distance and Friendship Links: Across-County Summary Statistics for the United States

	Share of friends living within:			*Share of US population living within:*		
	50 Miles	*100 Miles*	*200 Miles*	*50 Miles*	*100 Miles*	*200 Miles*
Mean	55.4%	62.8%	70.3%	1.3%	2.8%	6.6%
P5	38.1%	46.0%	54.2%	0.1%	0.3%	1.0%
P10	42.5%	49.6%	57.1%	0.1%	0.6%	2.1%
Median	55.4%	63.9%	71.6%	0.7%	2.1%	5.8%
P90	67.4%	74.8%	81.2%	3.2%	6.2%	15.0%
P95	70.3%	76.9%	83.2%	5.4%	9.2%	15.6%

Note: Table shows across-county summary statistics for the share of friends of a county's population living within a certain distance of that county as well as the share of the US population living within those distances. P5, P10, P90, and P95 are the 5th, 10th, 90th, and 95th percentiles, respectively. Counties are weighted by their populations.

The regressions in Table 2 offer a more systematic account of the relationship between geographic distance and social connectedness across county-pairs. The unit of observation is a county-pair. The dependent variable is the log of the Social Connectedness Index between the two counties. The log of the geographic distance between the counties is the explanatory variable in column 1. We include fixed effects for both counties, which controls for population levels and any other characteristics that vary at the county level. In this specification, geographic distance is able to explain a significant amount of the cross-county-pair variation in social connectedness. The estimated elasticity of social connectedness to geographic distance suggests that a 10 percent increase in the distance between two counties is associated with a 14.8 percent decline in the number of friendship links between those counties. Similar to gravity equations estimated in the trade literature, this estimates the equilibrium relationship between geographic distance and social connectedness, not necessarily the causal effect of one on the other.

In column 2, we include an additional control indicating whether both counties are within the same state. The social connectedness of a county is often strongest with other counties within the same state, even compared to nearby counties in other states. This finding is not the result of non-log linearities in the distance relationship, and it can be found for both border counties and nonborder counties (as we discuss further in the Appendix). Why social connectedness varies so strongly at state borders, and the extent to which this is driven by institutional, social, or economic factors, is an interesting avenue for future research. Possible explanations include the importance of common state-level identities or the role of state universities as meeting places for residents from the same state.

In columns 3 and 4, we restrict the sample to county-pairs that are more and less than 200 miles apart, respectively. In the sample of county-pairs that are less than

200 miles apart, the estimated elasticity between geographic distance and friendship links is –1.99. In the sample of county-pairs that are more than 200 miles apart, the magnitude of the elasticity falls by nearly half to –1.16. These findings suggest that while social connectedness is declining in geographic distance, the elasticity of this relationship is less negative as we include county-pairs that are progressively further apart. In turn, this pattern highlights that in the theoretical modeling of friendship links, the appropriate elasticity depends on the geographic distances studied. This finding may help to explain why previous estimates of the elasticity of friendship probability with respect to geographic distance vary so significantly across settings, including an estimate of –2 in a study of cell-phone communication networks in the United Kingdom (Lambiotte et al. 2008); an estimate of –1 among bloggers (Liben–Nowell, Novak, Kumar, Raghavan, and Tomkins 2005); and an estimate of –0.5 in location-based online social networks such as Brightkite, Foursquare, and Gowalla (Scellato, Noulas, Lambiotte, and Mascolo 2011).

A substantial literature has documented that individuals are more likely to be associated with other individuals of similar characteristics. Following Lazarsfeld and Merton (1954), this empirical regularity is referred to as “homophily.” Homophily has been documented for a large number of individual characteristics, including racial identity, gender, age, religion, and education, as well as intangible aspects such as attitudes and beliefs (for a comprehensive review of the literature, see McPherson, Smith-Lovin, and Cook 2001). Thus, in column 5 of Table 2 we add a number of variables measuring the similarity of counties on measures such as per capita income, education levels, and religiosity. We find that county pairs that are more similar on these dimensions have more friendship links. However, while the magnitude of the effect of these socioeconomic differences on social connectedness is potentially meaningful, adding them barely affects the coefficients on other explanatory variables or the R^2 relative to the specification in column 2.

Table 2 highlights that social connectedness drops off strongly at state borders. A related question is how closely the existing state borders resemble the borders that would form if we grouped together US counties to create communities with the aim of maximizing within-community social connectedness. There are a number of possible algorithms to facilitate such a grouping of counties. Here, we use a method called hierarchical agglomerative linkage clustering (which we describe further in the online Appendix).

Figure 2 shows the result when we use this algorithm to group the United States into 20 distinct communities. All resulting communities are spatially contiguous, which is a result of the strong dependence of social connectedness on geographic distance. In addition, and consistent with finding social connectedness to decline at state borders, many of the community borders line up with state borders. All of the West Coast States together with Nevada form one community. Similarly, all counties in states between New England and Pennsylvania are grouped into the same community. Another group of states is Florida, Georgia, and Alabama. However, some states are split into separate communities. The Texas panhandle is grouped with Oklahoma and Kansas, and Colorado’s Western Slope forms its own community.

Table 2
Determinants of Social Connectedness across County Pairs

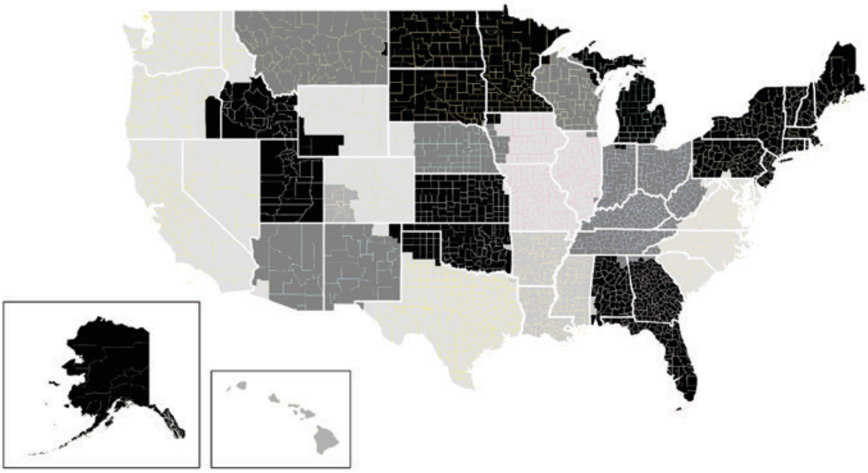
	Dependent Variable: Log(SCI)				
	(1)	(2)	(3)	(4)	(5)
log(Distance in Miles)	–1.483*** (0.065)	–1.287*** (0.061)	–1.160*** (0.059)	–1.988*** (0.043)	–1.214*** (0.055)
Same State		1.496*** (0.087)	1.271*** (0.083)	1.216*** (0.044)	1.496*** (0.085)
Δ Income ($1,000)					–0.006*** (0.001)
Δ Share Population White (%)					–0.012*** (0.001)
Δ Share Population No High School (%)					–0.012*** (0.002)
Δ 2008 Obama Vote Share (%)					–0.006*** (0.001)
Δ Share Population Religious (%)					–0.002*** (0.001)
County Fixed Effects	Y	Y	Y	Y	Y
Sample			>200 miles	<200 miles	
Number of observations	2,961,968	2,961,968	2,775,244	186,669	2,961,968
R^2	0.907	0.916	0.916	0.941	0.922

Note: Table shows results from a regression of the log of the Social Connectedness Index on a number of explanatory variables. The log of the geographic distance between the counties is the explanatory variable in column 1. In column 2, we include an additional control indicating whether both counties are within the same state. In columns 3 and 4, we restrict the sample to county-pairs that are more and less than 200 miles apart, respectively. The unit of observation is a county-pair. Standard errors are given in parentheses. The online Appendix (http://e-jep.org) provides more details on the data sources and exact specifications.
*, **, and *** indicate significance levels of $p < 0.1$, $p < 0.05$, and $p < 0.01$, respectively.

These findings suggest that it might be interesting to study the economics and politics of US "regions" as defined by joint social connectedness, rather than alternative groupings such as Census regions or divisions.

We have explored a number of additional correlates of friendship links across counties. For example, we document that the strength of social connections can be affected by physical obstacles such as large rivers and mountain ranges. We highlight that counties with military bases exhibit strong connections across the entirety of the United States, as do counties in North Dakota that have seen a recent shale oil boom and an associated significant in-migration. Counties with Native American reservations are strongly connected to one another. Similarly, areas with ski resorts in the Rocky Mountains and New England have high social connectedness. Counties in Florida with significant retiree populations are strongly connected to the Rust Belt and the Northeast. In addition, large cities in the Midwestern

Figure 2
Connected Communities within the United States—20 Units



Note: Figure shows US counties grouped together when we use hierarchical agglomerative linkage clustering to create 20 distinct groups of counties.

United States with significant African American populations, such as Milwaukee and Chicago, have strong links to the South around Mississippi and Alabama, consistent with friendship links persisting following the Great Migration of southern African Americans to northern cities. For more details on these patterns, see the online Appendix (http://e-jep.org). In general, many of these patterns of friendship connections are unsurprising, but it is new that such patterns can now be measured and documented in systematic national data.

Concentration of Social Networks and County Characteristics

The geographic concentrations of the friendship networks of different counties reveal a great deal of heterogeneity: for example, the earlier Table 1 shows that the 5th–95th percentile range across population-weighted counties in the share of friends living within 100 miles is 46.0 percent to 76.9 percent. Existing theoretical work suggests that the diversity of social networks is an important determinant of economic development; conversely, tightly clustered social ties can limit access to a broad range of social and economic opportunities (for example, Granovetter 1973). However, empirical studies of the relationship between the structure of social networks and economic outcomes of communities are rare. One exception is Eagle, Macy, and Claxton (2010), who use UK cellphone data to document that the diversity of individuals' social networks is correlated with regional economic well-being. In this section, we provide evidence that the geographic dispersion of friendship links across US counties is highly correlated with social and economic outcomes at the county level, such as average income, educational attainment, and social mobility.

If we define the concentration of a friendship network as the share of friends who live within 100 miles, then friendship networks in the South, the Midwest, and Appalachia are the most geographically concentrated. Counties in the Rocky Mountains have the smallest share of friends living within 100 miles, in large part because these areas are often less-densely populated. Among the western United States, Utah and inland California have the most geographically concentrated friendship networks. The online Appendix shows heat maps of this and other measures of the geographic concentration of friendship networks.

What are the effects of differentially structured social networks on county-level outcomes? As a first step toward answering this question, we correlate our measure of the concentration of friendship links with county-level characteristics. Figure 3 presents county-level binned scatterplots using the share of friends living within 100 miles and a number of socioeconomic outcomes. The overall message is that counties where people have more concentrated social networks tend to have worse socioeconomic outcomes along a number of dimensions: on average, they have lower income, lower education, higher teenage birth rate, lower life expectancy, less social capital, and less social mobility.

These correlations cannot be interpreted as causal (although the online Appendix discusses a number of causal mechanisms proposed by the literature that are consistent with our findings). Our goal here, as in the rest of the paper, is to document patterns that can guide future research investigating the causal effects of social network structure on socioeconomic outcomes, and to describe the Social Connectedness Index data that can help with such analyses. More generally, the strong correlation between social connectedness and socioeconomic outcomes suggests that controlling for the geographic concentration of social networks is important to minimize omitted variables bias across a number of research agendas that study economic and social outcomes at the county level.

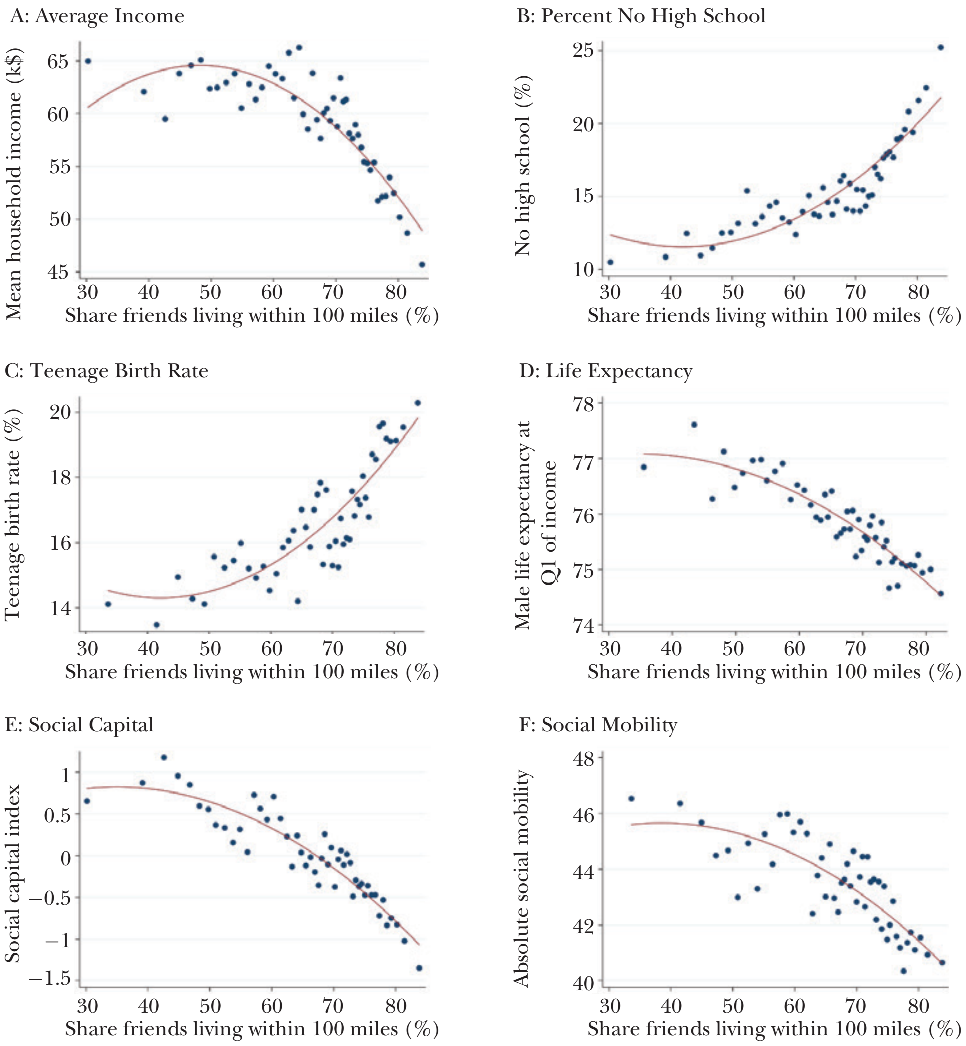
Social Connectedness and Cross-County Activity

Social connectedness between two regions may be related to other economic and social interactions between these regions. Indeed, we next document correlations between the number of friendship links and trade flows, patent citations, and migration patterns. As before, we illustrate some salient patterns in the data rather than providing full-fledged causal analyses. For each of the patterns documented below, the online Appendix (http://e-jep.org) provides more details on the variables, data construction, specifications, and additional exploration.

Social Connectedness and Within-US Trade Flows

A well-established empirical result in the trade literature is that bilateral trade between two regions decreases with geographic distance, although the explanations

Figure 3
Network Concentration and County-Level Characteristics



Notes: Panels show binned scatterplots with counties as the unit of observation. To generate each binned scatterplot, we group the x-axis variable into 50 equal-sized bins. We then compute the mean of the x-axis and y-axis variables within each bin and create a scatterplot of these 50 data points. The horizontal axes measure the share of friends of the county population that live within 100 miles. On the vertical axes are a number of county-level measures of socioeconomic outcomes: the mean county income in Panel A; the share of the population with no high school degree in Panel B; the teenage birth rate as provided by Chetty, Hendren, Kline, and Saez (2014) in Panel C; the life expectancy of males in the first quarter of the national income distribution from Chetty et al. (2016) in Panel D; the measure of social capital in 2009 as defined by Rupasingha, Goetz, and Freshwater (2006) in Panel E; and the absolute measure of social mobility from Chetty et al. (2014) in Panel F. The red line shows the fit of a quadratic regression. The online Appendix (http://e-jep.org) provides more details.

for this finding are still being debated (for a review, see Anderson and van Wincoop 2004). Many studies have highlighted that the distance effect is too large to be fully explained by trade costs alone, and that geographic distance might serve as a proxy for other trade frictions such as cultural differences, lack of familiarity, or information asymmetries. Social connections may alleviate the trade costs associated with these factors, and some empirical work has examined the causal effect of stronger social networks on trade (Rauch 1999; Combes, Lafourcade, and Mayer 2005; Cohen, Gurun, and Malloy 2012; Burchardi and Hassan 2013; Chaney 2014, 2016). However, much of this literature has struggled to measure the social connectedness between trading partners, and thus had to rely on indirect proxies, such as the ethnic composition of regions or past migration patterns.

The Social Connectedness Index data allow us to examine directly the empirical relationship between trade flows and social connectedness at the state level. Panel A of Table 3 shows some results. For the dependent variable, we measure interstate trading volumes using data from the Commodity Flow Survey. We focus on data from 2012, the latest year with comprehensively available data. Specifically, the dependent variable captures the log of the value of trade in 2012 between origination state *i* and destination state *j*.

For our main explanatory variables, we use the log of geographic distance between states *i* and *j*, as well as the log of the Social Connectedness Index between states *i* and *j* (constructed from a weighted average of county-level SCI measures). We also include fixed effects for each state, dummy variables for own-state flows, and dummy variables if the states are adjacent to each other.

We observe two main patterns. First, social connectedness is strongly correlated with state–state trade flows, even after controlling for geographic distance. The magnitude of the elasticity of trade with social connectedness is large and statistically significant.[3] In fact, when comparing them across columns 1 and 2, it appears as if social connectedness can explain marginally more of the variation in state–state trade flows than geographic distance.

Second, controlling for social connectedness significantly reduces the estimated distance elasticities of trade. A comparison of columns 1 and 3 shows that the distance elasticity of trade halves in magnitude after controlling for social connectedness. In column 4, we further control for differences across the states in GDP per capita, unemployment rates, sectoral composition, union share, and population density. The addition of these further controls has essentially no effect on the estimated elasticity between social connectedness and trade.

The observed reduction in the distance elasticities of trade, after controlling for social connectedness, is consistent with theories described above which suggest that geographic distance might be proxying for other factors affecting trade between

[3] In the online Appendix, we explore these patterns across industries. We find that the magnitude of the elasticity of trade flows with respect to friendship links rises with the share of high-skilled workers in the sector and is not affected by the share of labor compensation in total costs.

Table 3

Social Connectedness and Across-Region Economic Interactions

	(1)	(2)	(3)	(4)
Panel A: Dependent Variable: log(State-Level Trade Flows)				
log(Distance)	–1.057***		–0.531***	–0.533***
	(0.071)		(0.084)	(0.085)
log(SCI)		0.999***	0.643***	0.637***
		(0.051)	(0.071)	(0.060)
State Fixed Effects	Y	Y	Y	Y
Other State Differences	N	N	N	Y
Observations	2,219	2,220	2,219	2,219
R^2	0.912	0.918	0.926	0.930
Panel B: Dependent Variable: Indicator for Patent Citation				
log(Distance)	–0.048***		–0.011**	–0.021**
	(0.002)		(0.005)	(0.009)
log(SCI)		0.063***	0.049***	0.066***
		(0.003)	(0.006)	(0.012)
Technological Category + County Fixed Effects	Y	Y	Y	Y
Cited + Issued Patent Fixed Effects, Other County Differences	N	N	N	Y
Observations	2,171,754	2,171,754	2,171,754	2,168,285
R^2	0.056	0.059	0.059	0.101
Panel C: Dependent Variable: log(County-Level Migration)				
log(Distance)	–0.973***		0.023	0.031
	(0.048)		(0.021)	(0.021)
log(SCI)		1.134***	1.148***	1.159***
		(0.019)	(0.024)	(0.024)
County Fixed Effects	Y	Y	Y	Y
Other County Differences	N	N	N	Y
Observations	25,305	25,305	25,305	25,287
R^2	0.610	0.893	0.893	0.893

Note: Table shows the relationship between bilateral economic activity across geographic units and the geographic distance and social connectedness between these units. "SCI" stands for Social Connectedness Index. In Panel A, the unit of observation is a state-pair, and the dependent variable is the log of the value of 2012 trade flows between the states. All specifications include state fixed effects, dummies for own state, and dummies for neighboring states; column 4 also controls for differences across states on important socioeconomic indicators. In Panel B, the unit of observation is a patent-pair. The dependent variable is an indicator of whether patent *i* cites patent *j*. All specifications control for the county and technology category fixed effects, and column 4 also controls for patent fixed effects and other differences across the counties of the patents on important socioeconomic indicators. In Panel C, the unit of observation is a county pair, and the dependent variable is the log of across-county migration between 2013 and 2014. All specifications control for county fixed effects, and column 4 also controls for other differences across counties on important socioeconomic indicators. Standard errors are given in parentheses. The online Appendix (http://e-jep.org) provides more details on the data sources and exact specifications.
*, **, and *** indicate significance levels of $p < 0.1$, $p < 0.05$, and $p < 0.01$, respectively.

states. Further investigating the causal role of social connectedness in facilitating trade flows might therefore be a useful avenue for future research.

Social Connectedness and Patent Citations

In many models of endogenous growth, knowledge spillovers among individuals or firms are an important driver of productivity and economic growth (Romer 1986; Lucas 1988; Aghion and Howitt 1992). Social connectedness might therefore have important effects on economic activity, by facilitating the diffusion of knowledge and ideas through society.[4] However, testing these theories is challenging, because both knowledge spillovers and the degree of social connectedness are hard to measure. To overcome these challenges, a large empirical literature has relied on patent citations as a measure of knowledge spillovers (Jaffe, Trajtenberg, and Henderson 1993; Thompson and Fox-Kean 2005). By studying the geographic distances between the locations where the issued patents and patent citations occur, these papers conclude that knowledge spillovers are highly localized. In turn, this finding is often interpreted as evidence for the importance of social interactions, which are more likely to happen at shorter distances. Other attempts to measure social connectedness have tried to proxy for an inventor's peer group based on characteristics such as common ethnicity (Agrawal, Kapur, and McHale 2008).

The Social Connectedness Index has the potential to provide more direct evidence for the role of social connectedness in facilitating knowledge spillovers. We obtain data containing information on all patents granted by the US Patent and Trademark Office in the years 2002–2014, and the location of the company or institution from which the patent originated. If the company or institution is not available, then the patent is assigned to the location of the first inventor with an available location (as in Berkes and Gaetani 2017). The patents cover 107 different technological classes, defined based on the International Patent Classification. For each granted patent, we observe all other patents that it cites.

We follow the approach in the existing literature to explore the relationship between social connectedness and patent citations (for example, Jaffe, Trajtenberg, and Henderson 1993). This approach matches each "citing patent" with a "non-citing patent" issued at the same time and in the same technological class to serve as a control, as we will explain below. Knowledge spillovers are then measured as the extent to which the citation probability increases with the social connectedness of the geographies associated with the patents, after controlling for the patent's technological class and the geographic distance between the geographies. The literature has argued that this approach can help to separate knowledge spillovers

[4] For examples, see Jovanovic and Rob (1989), Kortum (1997), Benhabib and Spiegel (2005), Alvarez, Buera, and Lucas (2008), Comin and Hobijn (2010), Comin, Dmitriev, and Rossi-Hansberg (2012), Fogli and Veldkamp (2012), and Buera and Oberfield (2016). Social networks can also affect the exposure of the region to new ideas and thus how quickly the region adopts a new idea (for instance, Glaeser 1999; Black and Henderson 1999; Moretti 2012).

from correlations that might be induced by patterns in the geographic location of technologically related activities across regions that are connected through social networks.

To implement this approach, for each US patent granted in 2014, we create an observation for every patent cited by the 2014 patent, so that the unit of observation is a patent–citation pair. For example, if a particular 2014 patent cites 10 other patents, this will generate 10 patent–citation pairs. We then construct a control observation for each of these patent–citation pairs. In particular, for each 2014 patent *A* that cites a previous patent *B*, we randomly select another 2014 patent *C* that is in the same technology class as patent *A*, but that does not cite patent *B*. We focus on patent classes with at least 1,000 patents issued in 2014, to ensure that there is a sufficient sample to select the control patents randomly.

Panel B of Table 3 shows results from our analysis. The dependent variable in the regressions equals one if an issued patent *i* cites patent *j*, and zero otherwise. The first two rows show the coefficients on the log of geographic distance and the log of the Social Connectedness Index between the counties of the issued and cited patents. We include fixed effects for the technology classes and for the counties of patents *i* and *j*.

Comparing columns 1 and 2, social connectedness explains marginally more of the variation in the probability of a patent citation than geographic distance, as the R^2 in column 2 is higher. In terms of economic magnitudes, the probability of a patent citation is 6.3 percentage points higher when the social connectedness between the counties of the issued and cited patents doubles.

In column 3, we jointly estimate the relationship of geographic distance and social connectedness with the probability of a patent citation. The effect of doubling social connectedness on the probability of citation remains significant and large, at 4.9 percent, even after controlling for geographic distance. In comparison, the effect of doubling geographic distance on the probability of citations falls from –4.8 to –1.1 percent.

In column 4, we also control for a host of across-county differences on important socioeconomic indicators: 2008 vote share of Obama, mean income, share of population without a high school degree, share of population that is white, share of population that is religious, and share of workforce employed in manufacturing. We also add fixed effects for the cited and the issued patents. If anything, the estimated relationship between social connectedness and patent citation increases somewhat as a result of these further controls.

This finding suggests that the relationship between geographic distance and the probability of patent citation, viewed in isolation, may be partially capturing effects of information flows associated with social connectedness. More generally, our results suggest a significant correlation between social connectedness and knowledge spillovers, innovation, and, ultimately, economic growth. These findings highlight the potential of the Social Connectedness Index data to help uncover possible causal relationships behind these correlations.

Social Connectedness and Migration

Understanding the factors driving migration patterns is important. For example, within-US migration is one mechanism for equilibrating the US labor market following regional shocks (Blanchard and Katz 1992). An existing literature has documented that social networks can play an important role in facilitating migration by providing information as well as social and economic support (for a review, see Munshi 2016). While a lot of the research has focused on international migration (for example, Moretti 1999), similar forces might be at work in explaining within-US migration.

We find that the Social Connectedness Index has significant explanatory power for migration between regions, beyond what is predicted by geographic distance. Panel C of Table 3 shows some results. The dependent variable captures the log of total migration between counties *i* and *j* between 2013 and 2014, as measured by the Statistics of Income (SOI) Tax Stats Migration Data provided by the IRS. The key explanatory variables are the log of geographic distance between those counties and the log of the Social Connectedness Index. We also include fixed effects for each county, which allows us to control for the size of its population and other county-level characteristics that might affect the degree of migration.

In column 1 of Table 3, Panel C, we do not include the social connectedness variable. The estimated elasticity of migration to geographic distance is close to –1. In column 2, we find that the elasticity of migration to social connectedness is slightly larger than 1, with a somewhat higher R^2 than in column 1. In other words, the Social Connectedness Index can explain a larger part of the variation of the migration flows across county-pairs than geographic distance can. In column 3, we control for both the geographic distance and social connectedness between counties. We find that geographic distance adds no additional predictive power compared with column 2. This finding suggests that much of the estimated effect of distance on migration might be coming from the relationship between distance and social connectedness, and that distance by itself has no additional explanatory power for migration. Column 4 shows that these conclusions are robust to further controlling for other differences across counties on important socioeconomic indicators.

Overall, our results are consistent with stories in which individuals are more likely to move to counties where they already have friends. Such a mechanism could, for example, result in larger cities attracting even more new movers and thereby help explain the very right-tailed city size distribution (Gabaix 1999). Exploring the causal mechanisms behind the observed relationship between social connectedness and migration thus provides an exciting research agenda.

International Dimension of Social Connectedness of US Counties

US counties vary considerably in the share of social connections to individuals living outside of the United States. For the median county, 4 percent of all friendship links are to individuals living in foreign countries, but the 10–90 percentile range is 2.3 percent to 8.6 percent, and the 1–99 percentile range is 1.6 percent to 18.7 percent. Some of this variation is straightforward to explain. For example, areas close to the Mexican or the Canadian border have more international connections. Patterns of past immigration matter as well. For example, connections with Norway are particularly strong for those parts of the United States that saw major immigration from Norway in the late 19th and early 20th Centuries, like Wisconsin, Minnesota, and the Dakotas. Similarly, a number of counties in the northeastern United States have strong social connectedness to Italy. For heat maps of social connectedness to these and other countries, see the online Appendix available with this paper at http://e-jep.org.

The first three columns in Table 4 illustrate the extent to which past migration from a particular country is correlated with the strength of today's social connectedness of a US county with that country. In these columns, the dependent variable is the Social Connectedness Index between each county and foreign country. For the explanatory variables, geographic distance is measured between each county and the capital city of each foreign country. We use two measures of past migration: the number of residents who claim their primary ancestry as being from a given foreign country and the number of residents in each county who were born in a specific foreign country. The first measure is broader and can, for instance, include US-born individuals with immigrant parents or grandparents. All variables are measured in logs. We also include fixed effects for each county and foreign country.

The first column shows the correlation between geographic distance and international social connectedness: a 1 percent increase in the geographic distance is associated with a 1.2 percent decline in social connectedness. Interestingly, this elasticity is nearly identical to the elasticity of friendship links to geographic distance estimated for the United States for distances greater than 200 miles. The second column shows that a 1 percent increase in the number of residents with ancestry from a given foreign country correlates with an increase in social connections to that country by about one-third of a percent. In column 3, we obtain similar estimates for our second measure of past migration. Across columns 2 and 3, controlling for past migration reduces the estimated effect of geographic distance on social connectedness by between one-third and one-half.

In other regressions presented in the online Appendix, we find that the effect of past migration on today's social connections is stronger for countries from which immigration to the United Sates occurred more recently, such as Mexico or the Philippines, compared to countries from which immigration peaked earlier, such as Germany or Ireland. For example, the coefficient on a regression like that in column 2 is about 0.13 for counties with immigration waves that peaked pre-1900 or between 1900 and 1930, but more than twice as high for waves that peaked between 1930 and 1990 or for waves that have not yet peaked.

We also sought to estimate the relationship between social connectedness and international trade. Again, we used state-level data on social connectedness

Table 4
Social Connectedness, Ancestry, and International Trade

	log(SCI)			log (Exports + 1)	log (Imports + 1)
	(1)	(2)	(3)	(4)	(5)
log(Distance)	−1.159*** (0.258)	−0.690*** (0.162)	−0.493*** (0.174)	−2.092*** (0.391)	−1.627*** (0.378)
log(Ancestry in Foreign Country)		0.341*** (0.022)			
log(Born in Foreign Country)			0.367*** (0.033)		
log(SCI)				0.597*** (0.139)	0.470*** (0.103)
Fixed Effects	Y	Y	Y	Y	Y
Observations	33,146	33,146	16,527	11,015	11,014
R^2	0.908	0.936	0.943	0.770	0.770
Number of Countries	105	105	52	216	216

Note: The table explores the international dimension of social connectedness. In columns 1 to 3, we explore how past migration patterns and geographic distance are correlated with international social connectedness. The unit of observation is a US county–foreign country pair. Each specification also includes fixed effects for the US state and the foreign country, and the dependent variable is the log of the Social Connectedness Index between those units. In columns 4 and 5, we explore how today's international trading activity is correlated with social connectedness. The unit of observation is a US state–foreign country pair. Standard errors are given in parentheses. The online Appendix (http://e-jep.org) provides more details on the data sources and exact specifications.
*, **, and *** indicate significance levels of $p < 0.1$, $p < 0.05$, and $p < 0.01$, respectively.

(by combining the counties of a given state into a population-weighted average), because data on international trade is only available at the state level. Adjusting for geographic distance, (in a specification similar to Table 3, Panel B, column 3), we find that a state with 10 percent higher social connectedness to a given foreign country on average imports 4.7 percent more from this country and exports 6.0 percent more to this country. These findings are highly consistent with our earlier estimates on within-US trade. In the online Appendix for this paper, we provide additional details on these variables and alternative specifications.

Conclusion

We use data from the global online social networking site Facebook to construct the Social Connectedness Index (SCI). These data provide a new and comprehensive measure of social connectedness between US county pairs, as well as between US counties and foreign countries. The SCI should allow researchers to overcome some of the measurement challenges that have held back empirical research on the role of social interactions in finance, economics, and the broader social sciences. To illustrate this point, we show how the SCI data can be used to better understand the geographic dimensions of real-world social networks, as well as to document that social connectedness correlates strongly with social and economic activity across regions. While these correlations should not be seen as identifying causal relationships, they provide starting points for investigating a variety of important questions.

A number of recent studies have used data from online social networks, in most cases by including coauthors from Facebook or other social networking services. For example, Gee, Jones, and Burke (2017) and Gee, Jones, Fariss, Burke, and Fowler (2017) use de-identified microdata from Facebook to analyze the role of social networks in the job-finding process. These researchers were able to assess the relative importance of strong and weak ties in helping job seekers find new employment. Social network data from Facebook have also been used to study a range of other topics: the relationship between the size of friendship networks and mortality (Hobbs, Burke, Christakis, and Fowler 2016); the structure of social networks in immigrant communities in the United States (Herdağdelen, State, Adamic, and Mason 2016); the evolution of information cascades (Cheng, Adamic, Kleinberg, and Leskovec 2016); and the effects of social influence and social advertising (Bakshy, Eckles, Yan, and Rosenn 2012). Other researchers have studied the effects of online social networks themselves. For example, Bakshy, Messing, and Adamic (2015) study how online networks influence exposure to perspectives that cut across ideological lines. In our own work, we have used social network data from Facebook to document that social interactions influence people's perceptions of local housing markets as well as their real estate investment decisions and mortgage leverage choices (Bailey, Cao, Kuchler, and Stroebel forthcoming; Bailey, Dávila, Kuchler, and Stroebel 2017). We have also explored the role of peer effects in product adoption decisions (Bailey, Kuchler, Stroebel, and Wong 2018), and are working with other coauthors to better understand the role of social connectedness in facilitating social mobility.

For many researchers, it should prove a considerable advantage that the Social Connectedness Index is now more broadly available. In addition to the topics that we have explored in this paper, here are five other examples of policy and research questions that we hope will be pursued with the SCI data.

First, many contagious illnesses and diseases, such as the flu or tuberculosis, spread through human contact. Combined with localized data on the prevalence of the flu, data on social connectedness might allow researchers and public health officials to better predict where to expect future outbreaks of the flu (Cauchemez et al. 2011; Christakis and Fowler 2010).

Second, the Social Connectedness Index data could also be used to track whether measures of sentiment—for example, those tracked by the Michigan Survey of Consumers or through geo-coded Twitter feeds—spread along social networks.

Third, sociolinguistic research has argued that social networks are an important force determining how languages evolve over time (for example, Milroy 1987). The Social Connectedness Index data would allow researchers to study the extent to which linguistic development in the United States is associated with patterns of social connectedness.

Fourth, the relationships between transportation networks and social connectedness may prove interesting. For example, significant social connectedness between two regions might be a strong indicator that providing transportation infrastructure between these regions, such as direct airline routes, is profitable. Using the Social Connectedness Index as a measure of the potential demand for various routes could address some of the identification issues in the literature analyzing airline scheduling in operations research and industrial organization. Moreover, increased transportation links might also have a causal effect on social connectedness. One approach using the SCI data is to compare the social connectedness of two counties that happen to lie on the straight line between two major cities, and which are therefore connected by a highway, to the connectedness of two similar counties that do not lie on the straight line between major cities (see Bailey et al. 2018).

Finally, the SCI might prove useful in testing theoretical models of network formation (Jackson 2014). Specifically, in models of geographic strategic network formation models, the costs of network formation are directly related to distance (for example, Johnson and Gilles 2000). Using data from the National Longitudinal Survey of Adolescent Health on close friends of individuals, Patacchini, Picard, and Zenou (2015) show that students living in central locations have higher levels of social interactions. Our estimates of the elasticities of friendship links with respect to distance often map directly into the parameters of these models and can be used to parameterize them.

While we hope that the county-level Social Connectedness Index will prove useful to researchers, it is of course only one aspect of the vast wealth of data on networks being created by online social networking services. As these data become available in various forms, the modeling and analysis of social networks will advance substantially.

References

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth through Creative Destruction." *Econometrica* 60(2): 323–51.

Agrawal, Ajay, Devesh Kapur, and John McHale. 2008. "How Do Spatial and Social Proximity Influence Knowledge Flows? Evidence from Patent Data." *Journal of Urban Economics* 64(2): 258–69.

Alvarez, Fernando E., Francisco J. Buera, and Robert E. Lucas, Jr. 2008. "Models of Idea Flows." NBER Working Paper 14135.

Anderson, James E. and Eric van Wincoop. 2004. "Trade Costs." *Journal of Economic Literature* 42(3): 691–751.

Bailey, Michael, Ruiqing Cao, Theresa Kuchler, and Johannes Stroebel. Forthcoming. "The Economic Effects of Social Networks: Evidence from the Housing Market." *Journal of Political Economy.*

Bailey, Michael, Eduardo Dávila, Theresa Kuchler, and Johannes Stroebel. 2017. "House Price Beliefs and Mortgage Leverage Choice." NBER Working Paper 24091.

Bailey, Michael, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2018. "Peer Effects in

Product Adoption." Unpublished paper.

Bakshy, Eytan, Dean Eckles, Rong Yan, and Itamar Rosenn. 2012. "Social Influence in Social Advertising: Evidence from Field Experiments." In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*, pp. 146–61. New York, NY: ACM.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348(6239): 1130–32.

Benhabib, Jess, and Mark M. Spiegel. 2005. "Human Capital and Technology Diffusion." Chap. 13 in *Handbook of Economic Growth*, vol. 1A, edited by Philippe Aghion and Steven N. Durlauf, 935–966. Elsevier.

Berkes, Enrico, and Ruben Gaetani. 2017. "The Geography of Unconventional Innovation." https://cpb-us-east-1-juc1ugur1qwqqqo4.stackpathdns.com/sites.northwestern.edu/dist/4/638/files/2017/06/Berkes_Gaetani_Submission_June_2017-2ao3fck.pdf.

Black, Duncan, and Vernon Henderson. 1999. "A Theory of Urban Growth." *Journal of Political Economy* 107(2): 252–84.

Blanchard, Olivier Jean, and Lawrence F. Katz. 1992. "Regional Evolutions." *Brookings Papers on Economic Activity* no. 1, pp. 1–75.

Buera, Francisco J., and Ezra Oberfield. 2016. "The Global Diffusion of Ideas." NBER Working Paper 21844.

Burchardi, Konrad B., and Tarek A. Hassan. 2013. "The Economic Impact of Social Ties: Evidence from German Reunification." *Quarterly Journal of Economics* 128(3): 1219–71.

Cauchemez, Simon, Achuyt Bhattarai, Tiffany L. Marchbanks, Ryan P. Fagan, Stephen Ostroff, Neil M. Ferguson, David Swerdlow, and the Pennsylvania H1N1 Working Group. 2011. "Role of Social Networks in Shaping Disease Transmission during a Community Outbreak of 2009 H1N1 Pandemic Influenza." *PNAS* 108(7): 2825–30.

Chaney, Thomas. 2014. "The Network Structure of International Trade." *American Economic Review* 104(11): 3600–34.

Chaney, Thomas. 2016. "Networks in International Trade." Chap. 28 in *Oxford Handbook of the Economics of Networks*, edited by Yann Bramoullé, Andrea Galeotti, and Brian Rogers. Oxford University Press.

Cheng, Justin, Lada A. Adamic, Jon M. Kleinberg, and Jure Leskovec. 2016. "Do Cascades Recur?" 671–681, International World Wide Web Conferences Steering Committee. arXiv:1602.01107 [cs.SI].

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *Quarterly Journal of Economics* 129(4): 1553–1623.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. "The Association between Income and Life Expectancy in the United States, 2001–2014." *JAMA* 315(16): 1750–66.

Christakis, Nicholas A., and James H. Fowler. 2010. "Social Network Sensors for Early Detection of Contagious Outbreaks." *PloS ONE* 5(9): e12948.

Cohen, Lauren, Umit G. Gurun, and Christopher J. Malloy. 2012. "Resident Networks and Firm Trade." NBER Working Paper 18312.

Combes, Pierre-Philippe, Miren Lafourcade, and Thierry Mayer. 2005. "The Trade-Creating Effects of Business and Social Networks: Evidence from France." *Journal of International Economics* 66(1): 1–29.

Comin, Diego A., Mikhail Dmitriev, and Esteban Rossi-Hansberg. 2012. "The Spatial Diffusion of Technology." NBER Working Paper 18534.

Comin, Diego, and Bart Hobijn. 2010. "An Exploration of Technology Diffusion." *American Economic Review* 100(5): 2031–59.

Duggan, Maeve, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. 2015. "Social Media Update 2014." Pew Research Center, January 9.

Eagle, Nathan, Michael Macy, and Rob Claxton. 2010. "Network Diversity and Economic Development." *Science,* May 21, 328(5981): 1029–31.

Facebook. 2017. "Facebook Form 10-Q, Quarter 4, 2017." https://s21.q4cdn.com/399680738/files/doc_financials/2017/Q4/Q4-2017-Earnings-Presentation.pdf.

Fogli, Alessandra, and Laura Veldkamp. 2012. "Germs, Social Networks and Growth." NBER Working Paper 18470.

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

Gee, Laura K., Jason Jones, and Moira Burke. 2017. "Social Networks and Labor Markets: How Strong Ties Relate to Job Finding on Facebook's Social Network." *Journal of Labor Economics* 35(2): 485–518.

Gee, Laura K., Jason J. Jones, Christopher J. Fariss, Moira Burke, and James H. Fowler. 2017. "The Paradox of Weak Ties in 55 Countries." *Journal of Economic Behavior & Organization* 133: 362–72.

Gilbert, Eric, and Karrie Karahalios. 2009. "Predicting Tie Strength with Social Media." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 211–20. ACM.

Glaeser, Edward. 1999. "Learning in Cities."

Journal of Urban Economics 46(2): 254–77.

Granovetter, Mark S. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78(6): 1360–80.

Hampton, Keith, Lauren Sessions Goulet, Lee Rainie, and Kristen Purcell. 2011. "Social Networking Sites and Our Lives." Pew Internet and American Life Project, Washington, DC.

Herdağdelen, Amaç, Bogdan State, Lada Adamic, and Winter Mason. 2016. "The Social Ties of Immigrant Communities in the United States." *Proceedings of the 8th ACM Conference on Web Science,* 78–84.

Hobbs, William R., Moira Burke, Nicholas A. Christakis, and James H. Fowler. 2016. "Online Social Integration is Associated with Reduced Mortality Risk." *PNAS* 113(46): 12980–984.

Jackson, Matthew O. 2014. "Networks in the Understanding of Economic Behaviors." *Journal of Economic Perspectives* 28(4): 3–22.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Quarterly Journal of Economics* 108(3): 577–98.

Johnson, Cathleen, and Robert Gilles. 2000. "Spatial Social Networks." *Review of Economic Design* 5(3): 273–99.

Jones, Jason J., Jaime E. Settle, Robert M. Bond, Christopher J. Fariss, Cameron Marlow, and James H. Fowler. 2013. "Inferring Tie Strength from Online Directed Behavior." *PloS ONE* 8(1): e52168.

Jovanovic, Boyan, and Rafael Rob. 1989. "The Growth and Diffusion of Knowledge." *Review of Economic Studies* 56(4): 569–82.

Kortum, Samuel S. 1997. "Research, Patenting, and Technological Change." *Econometrica* 65(6): 1389–1419.

Lambiotte, Renaud, Vincent D. Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. 2008. "Geographical Dispersal of Mobile Communication Networks." *Physica A: Statistical Mechanics and Its Applications* 387(21): 5317–25.

Lazarsfeld, P., and R. K. Merton. 1954. "Friendship as a Social Process: A Substantive and Methodological Analysis." In *Freedom and Control in Modern Society,* edited by Morroe Berger, Theodore Abel, and Charles H. Page. New York: Van Nostrand.

Liben-Nowell, David, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2005. "Geographic Routing in Social Networks." *PNAS* 102(33): 11623–28.

Lucas, Robert E., Jr. 1988. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22(1): 3–42.

Marmaros, David, and Bruce Sacerdote. 2006. "How Do Friendships Form?" *Quarterly Journal of Economics* 121(1): 79–119.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415–44.

Milroy, Lesley. 1987. *Language and Social Networks.* 2nd edition. Wiley-Blackwell.

Moretti, Enrico. 1999. "Social Networks and Migrations: Italy 1876–1913." *International Migration Review* 33(3): 640–57.

Moretti, Enrico. 2012. *The New Geography of Jobs.* Houghton Mifflin Harcourt.

Munshi, Kaivan. 2016. "Community Networks and Migration." Chap. 23 in *The Oxford Handbook of the Economics of Networks,* edited by Yann Bramoullé, Andrea Galeotti, and Brian Rogers. Oxford University Press.

Patacchini, Eleonora, Pierre M. Picard, and Yves Zenou. 2015. "Urban Social Structure, Social Capital and Spatial Proximity." CEPR Discussion Papers no. DP10501, Center for Economic Policy Research, March.

Rauch, James E. 1999. "Networks versus Markets in International Trade." *Journal of International Economics* 48(1): 7–35.

Romer, Paul M. 1986. "Increasing Returns and Long-Run Growth." *Journal of Political Economy* 94(5): 1002–37.

Rupasingha, Anil, Stephan J. Goetz, and David Freshwater. 2006. "The Production of Social Capital in US Counties." *Journal of Socio-Economics* 35(1: Essays on Behavioral Economics): 83–101.

Scellato, Salvatore, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. 2011. "Socio-Spatial Properties of Online Location-Based Social Networks." *Proceedings of the Fifth International Conference on Weblogs and Social Media,* held in Barcelona, Catalonia, Spain, July 17–21. AAAI Digital Library.

Thompson, Peter, and Melanie Fox-Kean. 2005. "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." *American Economic Review* 95(1): 450–60.

Verbrugge, Lois M. 1983. "A Research Note on Adult Friendship Contact: A Dyadic Perspective." *Social Forces* 62(1): 78–83.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley Press.

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 281–288

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., St. Paul, MN 55105.

Smorgasbord

The *Annual Review of Public Health* has published a pro-and-con on whether e-cigarettes are a plausible method of reducing harms from tobacco use. On one side, David B. Abrams, Allison M. Glasser, Jennifer L. Pearson, Andrea C. Villanti, Lauren K. Collins, and Raymond S. Niaura have written "A Harm Minimization and Tobacco Control: Reframing Societal Views of Nicotine Use to Rapidly Save Lives." "A diverse class of alternative nicotine delivery systems (ANDS) has recently been developed that do not combust tobacco and are substantially less harmful than cigarettes. ANDS have the potential to disrupt the 120-year dominance of the cigarette and challenge the field on how the tobacco pandemic could be reversed if nicotine is decoupled from lethal inhaled smoke. ANDS may provide a means to

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.3.281 doi=10.1257/jep.32.3.281

compete with, and even replace, combusted cigarette use, saving more lives more rapidly than previously possible." Stanton A. Glantz and David W. Bareham offer a skeptical view in "E-Cigarettes: Use, Effects on Smoking, Risks, and Policy Implications" (pp. 215–35). "While e-cigarettes deliver lower levels of carcinogens than do conventional cigarettes, they still expose users to high levels of ultrafine particles and other toxins that may substantially increase cardiovascular and noncancer lung disease risks, which account for more than half of all smoking-caused deaths, at rates similar to conventional cigarettes. Moreover, rather than stimulating smokers to switch from conventional cigarettes to less dangerous e-cigarettes or quitting altogether, e-cigarettes are reducing smoking cessation rates and expanding the nicotine market by attracting youth." In the April 2018 issue, Abrams et al. are at pp. 193–213, https://www.annualreviews.org/doi/10.1146/annurev-publhealth-040617-013849, while Glantz and Bareham are at pp. 215–235, https://www.annualreviews.org/doi/10.1146/annurev-publhealth-040617-013757.

Marion Fourcade delivered a keynote address at the 2017 meetings of the Swiss Society of Economics and Statistics on the topic: "Economics: The View from Below." "As Robert Chernomas and Ian Hudson put it, 'economics has the awkward distinction of being both the most influential and the most reviled social science.' We might add: economics may be the most reviled social science precisely because it is the most influential. ... Unlike the other social scientific disciplines, economics comes with a promise: the promise to make money, the promise to save money, the promise to allocate money (a rare resource) in the most efficient manner. In other words, part of the authority of economists also comes from their association with whoever holds the purse strings. They navigate the most powerful parts of the world, where financial decisions are being made and where political and corporate leaders are being trained. And, I shall add, this association has become increasingly tight over the course of the twentieth century. Business schools, for instance, have gone from being intellectual backwaters staffed with practitioners to becoming scientific powerhouses filled with disciplinary social scientists (with economics PhDs being the largest group). The consequences of this prosperous social position are not trivial. Let us remember that money is not neutral. It changes people from within." *Swiss Journal of Economics and Statistics,* 2018, vol. 154, article 5, at https://sjes.springeropen.com/articles/10.1186/s41937-017-0019-2, or watch video of the presentation at https://sjes.springeropen.com/articles/10.1186/s41937-017-0019. The essay can be read in combination with Fourcade's arguments in "The Superiority of Economists," coauthored with Etienne Ollion and Yann Algan, in the Winter 2015 issue of this journal.

Pathways, published by the Stanford Center on Poverty & Inequality, offers nine short and readable essays by social scientists and a few politicians on "The Next Round of Welfare Reform." For example, Robert A. Moffitt and Stephanie Garlow (pp. 17–21) discuss "Did Welfare Reform Increase Employment and Reduce Poverty?" "*Did welfare reform reduce welfare recipiency?* The welfare rolls indeed plummeted under the influence of [the 1996] welfare reform. If anything, some of the early studies underestimated the causal effect of welfare reform itself (as against the effects of economic expansion). *Did it increase employment?* Although there remains some ambiguity on the relative importance of the EITC and welfare reform in accounting for changes in employment, it is clear that welfare reform played an important role. In the initial years after reform, many more women joined the labor force than even the reform's most ardent supporters had hoped. *Did it reduce poverty?* There are two sides to the answer to this question. It would appear that, while welfare reform assisted families with incomes close to the poverty threshold, it did less to help families in deep or extreme poverty. Under the current welfare regime, many single mothers are struggling to support their families without income or cash benefits. Even women who are willing to work often cannot find good-paying, steady employment." Winter 2018, https://inequality.stanford.edu/publications/pathway/next-round-welfare-reform.

Marie-Anne Valfort has written "LGBTI in OECD Countries: A Review." "This paper presents an overview of the socio-economic situation of lesbians, gay men, bisexuals, transgender and intersex people (LGBTI), primarily in OECD countries. After investigating the size of this population, the paper zooms in on attitudes toward LGBTI, LGBTI rights and perceived discrimination among LGBTI. It goes on to discuss the empirical strategies used to identify whether LGBTI fare worse than non-LGBTI and provides a systematic review of survey-based and experimental evidence on such an 'LGBTI penalty' and its causes. This exploration points to substantial hurdles for LGBTI. In particular, (i) low legal recognition of same-sex couples hampers partnership stability and children's well-being; (ii) LGBTI are bullied at school and suffer academically; (iii) LGBTI face hiring and wage discrimination; (iv) LGBTI show higher rates of physical and mental health problems, in particular due to social rejection." OECD Social, Employment and Migration Working Papers No. 198, June 22, 2017, https://www.oecd-ilibrary.org/social-issues-migration-health/lgbti-in-oecd-countries_d5d49711-en.

Stephen T. Anderson discusses idiosyncrasies of the helium market in "Economics, Helium, and the U.S. Federal Helium Reserve: Summary and Outlook." From the abstract: "In 2017, disruptions in the global supply of helium reminded consumers, distributors, and policy makers that the global helium supply chain lacks flexibility, and that attempts to increase production from the U.S. Federal Helium Reserve (the FHR) may not be able to compensate for the loss of one of the few major producers in the world. Issues with U.S. and global markets for helium include inelastic demand, economic availability of helium only as a byproduct, only 4–5 major producers, helium's propensity to escape earth's crust, an ongoing absence of storage facilities comparable to the FHR, and a lack of consequences for the venting of helium. The complex combination of these economic, physical, and regulatory issues is unique to helium, and determining helium's practical availability goes far beyond estimating the technically accessible volume of underground resources." *Natural Resources Research*, October 2018, vol. 27, no. 4, pp. 455–477, https://link.springer.com/content/pdf/10.1007%2Fs11053-017-9359-y.pdf.

More on Macroeconomics

This issue of JEP features a seven-paper symposium on macroeconomics. Want more? The *Oxford Review of Economic Policy* has devoted a special double issue to a symposium on the topic of "Rebuilding Macroeconomic Theory." Contributors include Olivier Blanchard, Paul Krugman, Joseph Stiglitz, Simon Wren-Lewis, and others. As one example, Ricardo Reis asks "Is Something Really Wrong with Macroeconomics?" "Imagine going to your doctor and asking her to forecast whether you will be alive 2 years from now. That would sound like a preposterous request to the physician, but perhaps having some actuarial mortality tables in her head, she would tell you the probability of death for someone of your age. For all but the older readers of this article, this will be well below 50 per cent. Yet, 1 year later, you have a heart attack and die. Should there be outrage at the state of medicine for missing the forecast, with such deadly consequences? One defence by the medical profession would be to say that their job is not to predict time of death. They are driven to understand what causes diseases, how to prevent them, how to treat them, and altogether how to lower the chances of mortality while trading this off against life quality and satisfaction. Shocks are by definition unexpected, they cannot be predicted … This argument applies, word for word, to economics once the word 'disease' is replaced by the words 'financial crisis'. ... Too many people all over the world are today being unexpectedly diagnosed with cancer, undergo enormously painful treatment, and recover to live for many more years. This is rightly hailed as a triumph of modern oncology, even if so much more remains to be done. After suffering the worst shock in many decades, the global economy's problems were diagnosed by economists, who designed policies to respond to them, and in the end we had a painful recession but no melt-down. Some, somehow, conclude that economics is at fault." Spring–Summer 2018, https://academic.oup.com/oxrep/issue/34/1-2 (subscription required).

Edward Nelson suggests "Seven Fallacies Concerning Milton Friedman's 'The Role of Monetary Policy.'" "Fallacy 1: 'The Role of Monetary Policy' was Friedman's first public statement of the natural rate hypothesis." "Fallacy 2: The Friedman-Phelps Phillips curve was already presented in Samuelson and Solow's (1960) analysis." "Fallacy 3: Friedman's specification of the Phillips curve was based on perfect competition and no nominal rigidities." "Fallacy 4: Friedman's (1968) account of monetary policy in the Great Depression contradicted the *Monetary History*'s version." "Fallacy 5: Friedman (1968) stated that a monetary expansion will keep the unemployment rate and the real interest rate below their natural rates for two decades." "Fallacy 6: The zero lower bound on nominal interest rates invalidates the natural rate hypothesis." "Fallacy 7: Friedman's (1968) treatment of an interest-rate peg was refuted by the rational expectations revolution." Federal Reserve System, Finance and Economics Discussion Series 2018-013, https://www.federalreserve.gov/econres/feds/files/2018013pap.pdf.

The European Union

In Barry Eichengreen's "Euro Malaise: From Remission to Cure," he diagnoses five main issues of the euro: "First, Europe has a financial-stability problem. As a result of bad management, bad supervision and badly designed regulation, euro-area banks became deeply entangled in the global financial crisis. … European regulators were then slow to clean up the post-meltdown mess, which goes a long way toward explaining why Europe's recovery has been so sluggish. Second, the euro area has a debt problem. ... Third (and relatedly), fiscal policy is a problem. The euro area has an elaborate set of fiscal rules that are honored mainly in the breach. … Although the rules in question specify sanctions and fines for violators, those fines have never once been levied in the eurozone's almost two decades of existence. Fourth, the euro area lacks an adequate financial fire brigade, a regional equivalent of the International Monetary Fund. ... Fifth, the euro area lacks the flexibility to adjust to what the economist Robert Mundell, the intellectual father of the euro, referred to as 'asymmetric disturbances.' There is no mechanism for eliminating the imbalances that arise when some member-states are booming while others are depressed, or when some members increase productivity more rapidly than others." *Milken Institute Review,* First Quarter 2018, pp. http://www.milkenreview.org/articles/euro-malaise-from-remission-to-cure.

Jacob Funk Kirkegaard and Adam S. Posen have edited a collection of five essays for the European Commission, published in *Lessons for EU Integration from US History*. They write in in their overview essay, "Realistic European Integration in Light of US Economic History": "It is not important whether the European Union is integrating more or less quickly than the United States did. Such abstract benchmarking misses all the important points about the nature and sequencing of integration as political processes. The many fundamental differences between the United States and the European Union prevent drawing too precise, let alone literal, a mapping from US economic development to Europe's path forward today. ... Rather than pointing towards the current state of US continental integration as the guide for the European Union, we analyze the US responses throughout history to economic and political challenges and to numerous domestic political constraints—some not unlike what Europe faces today. We believe that EU leaders should draw lessons from these US responses for how, how far, and how fast their aspirations for EMU should progress. Yet, it must be acknowledged that the United States solved most of its political and economic challenges through centralization and federal government institution building." January 2018, Peterson Institute for International Economics, https://piie.com/system/files/documents/kirkegaard-posen_ec-report2018-01.pdf.

Significance of *p*-Values

Ronald L. Wasserstein and Nicole A. Lazar begin their discussion of "The ASA's Statement on *p*-Values: Context, Process, and Purpose" with this anecdote: "In

February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum: Q: Why do so many colleges and grad schools teach $p = 0.05$? A: Because that's still what the scientific community and journal editors use. Q: Why do so many people still use $p = 0.05$? A: Because that's what they were taught in college or grad school." *American Statistician*, 2016, vol. 70, no. 2, pp. 129–132, https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Wze4ntJKjDc.

A group of 72 empirical researchers followed up with this call: "Redefine Statistical Significance: We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries." Daniel J. Benjamin et al., *Nature Human Behavior* January 2018, pp. 6–10, https://www.nature.com/articles/s41562-017-0189-z.pdf. One of the signatories, John P. A. Ioannidis, provides an overview in the *JAMA* article "Viewpoint: The Proposal to Lower *P* Value Thresholds to .005." He writes: "*P* values and accompanying methods of statistical significance testing are creating challenges in biomedical science and other disciplines. The vast majority (96%) of articles that report *P* values in the abstract, full text, or both include some values of .05 or less. However, many of the claims that these reports highlight are likely false. … The status quo is widely believed to be problematic, but how exactly to fix the problem is far more contentious. ... Another large coalition of 72 methodologists recently proposed a specific, simple move: lowering the routine *P* value threshold for claiming statistical significance from .05 to .005 for new discoveries. … *P* values are misinterpreted, overtrusted, and misused. ... Moving the *P* value threshold from .05 to .005 will shift about one-third of the statistically significant results of past biomedical literature to the category of just 'suggestive.'" *Journal of the American Medical Association*, March 22, 2018, pp. E1–E2, https://jamanetwork.com/journals/jama/article-abstract/2676503 (registration needed).

Interviews

David S. Price serves as interlocutor in "Interview: Jean Tirole." "[N]ew platforms have natural monopoly features, in that they exhibit large network externalities. … I use the Google search engine or Waze because there are many people using it, so the algorithms are built on more data and predict better. Network externalities tend to create monopolies or tight oligopolies. So we have to take that into account. Maybe not by breaking them up, because it's hard to break up such firms: Unlike for AT&T or power companies in the past, the technology changes very fast; besides, many of the services are built on data that are common to all services. But to keep the market contestable, we must prevent the tech giants from swallowing up their future competitors; easier said than done of course ... Bundling practices by the tech giants are also of concern. A startup that may become an efficient competitor to such firms generally enters within a market niche; it's very hard to enter all segments at the same time. Therefore, bundling may prevent efficient entrants from entering market segments and collectively challenging the incumbent on the overall technology. Another issue is that most platforms offer you a best price guarantee, also called a 'most favored nation' clause or a price parity clause. You as a consumer are guaranteed to get the lowest price on the platform, as required from the merchants. Sounds good, except that if all or most merchants are listed on the platform and the platform is guaranteed the lowest price, there is no incentive for you to look anywhere else; you have become a 'unique' customer, and so the platform can set large fees to the merchant to get access to you. Interestingly, due to price uniformity, these fees are paid by both platform and nonplatform users—so each platform succeeds in taxing its rivals! That can sometimes be quite problematic for competition." *Econ Focus*, Federal Reserve Bank of Richmond, Fourth Quarter 2017, pp. 22–27, https://www.richmondfed.org/publications/research/econ_focus/2017/q4/interview.

Constantinos Repapis interviews "Professor Julie Nelson on Feminist Economics." "[W]hen people hear 'Women are more risk-averse,' people tend to think of that as categorical—women over here, men over there. In my meta-analysis, I looked back at the statistical data on which this claim was based and the two distributions are almost entirely overlapping. There is at least 80%, sometimes 90 or 96% overlap between the men's and women's distributions. There may also be tiny, perhaps statistically significant differences in the means of the distributions, but men and women are really a lot more similar than different. Yet, if you read the titles of certain books or articles, you would be getting a big misperception. ... [T]o me, feminism is not treating women as second-class citizens, as there to help and entertain men. And then my more methodological work has been about the biases that have been built into economics by choosing only the masculine-associated parts of life and techniques and banishing the feminine-associated ones. In my own life, I'm quite comfortable in both economics and feminist camps. I find when I give talks I get interesting labels. When I talk to a group of relatively mainstream economists I'm a wild-eyed radical leftist feminist nutcase. But because I'm an economist, when I talk to a lot of gender and women's studies groups, and I don't talk about the evils of global corporate capitalism and I don't have a certain line that I take on the economy, I'm considered a right-wing apologist for capitalism. And I'm quite comfortable balancing those two." Goldsmiths Economics, January 17, 2017, http://www.economicsppf.com/julie-nelson.html. The interview with Nelson is one in an ongoing project at at Goldsmiths, University of London, run by Ivano Cardinale and Constantinos Repapis. At http://www.economicsppf.com/index.html, they also have posted video and transcripts of substantial interviews done in the last few years with five other prominent economists who, in different ways, would classify themselves as being out of the mainstream of the profession: Sheila Dow, Geoff Harcourt, Charles Goodhart, Tony Lawson, and Ha-Joon Chang.

Discussion Starters

Philip J. Cook and Kimberly D. Krawiec offer this topic for kicking-off a conversation: "If we pay football players, why not kidney donors?" "In the United States and most every other country (with the notable exception of Iran), kidney donation is permitted but financial compensation for donors is prohibited. … The ban on compensation may protect potential donors from the temptation of easing their financial situation by giving up a kidney, a choice they may regret in later years. But this regulation has dire consequences. … The official waiting list of Americans with renal failure is now approximately 100,000, with a typical wait time of five years or more. … If ethical concerns persuade thoughtful people that the 'right' answer is to ban compensation for kidney donation, then the same logic would suggest that compensation should also be banned for participation in violent sports. If the 'right' answer is to permit compensation for participation in violent sports, then compensation for kidney donation should also be permitted. We see no logical basis for the current combination of banning compensation for kidney donors while allowing compensation for football players and boxers." *Regulation*, Spring 2018, pp. 12–17, https://object.cato.org/sites/cato.org/files/serials/files/regulation/2018/3/regulation-v41n1-4.pdf.

"At $164 trillion—equivalent to 225 percent of global GDP—global debt continues to hit new record highs almost a decade after the collapse of Lehman Brothers. Compared with the previous peak in 2009, the world is now 12 percent of GDP deeper in debt, reflecting a pickup in both public and nonfinancial private sector debt after a short hiatus. … Only three countries (China, Japan, United States) account for more than half of global debt—significantly greater than their share of global output." The comment is from chapter 1, "Saving for a Rainy Day," in the April 2018 issue of the *IMF Fiscal Monitor*, http://www.imf.org/en/Publications/FM/Issues/2018/04/06/fiscal-monitor-april-2018.

David Schleicher discusses "Stuck! The Law and Economics of Residential Stagnation." "Leaving one's home in search of a better life is, perhaps, the most classic of all American stories. ... But today, the number of Americans who leave home for new opportunities is in decline. A series of studies shows that the interstate migration rate has fallen substantially since the 1980s. Americans now move less often than Canadians, and no more than Finns or Danes. ... First, fewer Americans are moving away from geographic areas of low economic opportunity. … Americans, especially those who are non-college educated, are choosing to stay in areas hit by negative economic shocks. There is a long history of localized shocks generating interstate mobility in the United States; today, however, economists at the International Monetary Fund note that 'following the same negative shock to labor demand, affected workers have more and more tended to either drop out of the labor force or remain unemployed instead of relocating.' Second, lower-skilled workers are not moving to high-wage cities and regions. Bankers and technologists continue to move from Mississippi or Arkansas to New York or Silicon Valley, but few janitors make similar moves, despite the higher nominal wages on offer in rich regions for all types of jobs. As a result, local economic booms no longer create boomtowns. … Inequality between states has become entrenched." Schleicher explores state, local, and federal policies that "have created substantial barriers to interstate mobility, particularly for lower-income Americans." *Yale Law Journal*, October 2017, vol. 127, no. 1, pp. 78–154, https://www.yalelawjournal.org/article/stuck-the-law-and-economics-of-residential-stagnation.

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary *JEP* for your e-reader, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $20.00, $30.00, or $40.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

Founded in 1885

The Journal of

Economic Perspectives

Summer 2018, Volume 32, Number 3

Symposia

Macroeconomics a Decade after the Great Recession

Incentives in the Workplace

Articles

Features

AMERICAN
ECONOMIC
ASSOCIATION