The Journal of Economic Perspectives

A journal of the
American Economic Association

Winter 2018

The Journal of
Economic Perspectives

A journal of the American Economic Association

Editor
Enrico Moretti, University of California at Berkeley

Coeditors
Gordon Hanson, University of California at San Diego
Mark Gertler, New York University

Associate Editors
Nicholas Bloom, Stanford University
Leah Boustan, Princeton University
Dora Costa, University of California at Los Angeles
Amy Finkelstein, Massachusetts Institute of Technology
Seema Jayachandran, Northwestern University
Guido Lorenzoni, Northwestern University
Valerie Ramey, University of California at San Diego
Fiona Scott Morton, Yale University
Betsey Stevenson, University of Michigan
Ebonya Washington, Yale University
Catherine Wolfram, University of California
Luigi Zingales, University of Chicago

Managing Editor
Timothy Taylor

Assistant Editor
Ann Norman

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 (®).
Copyright © 2018 by the American Economic Association; All Rights Reserved.
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), Winter 2018, Vol. 32, No. 1. 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 1 • Winter 2018

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

Advisory Board

Kristen Butcher, Wellesley College
Janet Currie, Princeton University
Claudia Goldin, Harvard University
Robert E. Hall, Stanford University
Trevon Logan, Ohio State University
Scott Page, University of Michigan
Eduardo Porter, *New York Times*
Paul Romer, World Bank
David Sappington, University of Florida
Elu von Thadden, University of Mannheim

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 3–30

The Economic Implications of Housing Supply

Edward Glaeser and Joseph Gyourko

For most of US history, local economic booms were matched by local building booms. Into the 1960s, building was lightly regulated almost everywhere. Much housing was built in all high demand areas, including coastal California and New York City. However, between the 1960s and the 1990s, it became far more difficult to build in some areas with strong economic growth, especially those along the coasts. For example, there were 13,000 new housing units permitted in Manhattan in the single year of 1960 alone, which is nearly two-thirds of the 21,000 new units permitted throughout the decade of the 1990s (Glaeser, Gyourko, and Saks 2005). Higher economic productivity in the San Francisco Bay area, with its extensive restrictions on land use and building, now leads primarily to higher housing prices, rather than more homes and more workers (Ganong and Shoag 2013).

In this essay, we review the basic economics of housing supply and the functioning of US housing markets to better understand the distribution of home prices, household wealth, and the spatial distribution of people across markets. We employ a cost-based approach to gauge whether a housing market is delivering appropriately priced units. Specifically, we investigate whether market prices (roughly) equal the costs of producing the housing unit. If so, the market is well-functioning in the

■ *Edward Glaeser is the Fred and Eleanor Glimp Professor of Economics, Harvard University, Cambridge, Massachusetts. Joseph Gyourko is the Martin Bucksbaum Professor of Real Estate, Finance and Business Economics and Public Policy, the Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania. Their email addresses are eglaeser@harvard.edu and gyourko@wharton.upenn.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.1.3

doi=10.1257/jep.32.1.3

sense that it efficiently delivers housing units at their production cost. Of course, poorer households still may have very high housing cost burdens that society may wish to address via transfers. But if housing prices are above this cost in a given area, then the housing market is not functioning well—and housing is too expensive for all households in the market, not just for poorer ones.[1] The gap between price and production cost can be understood as a regulatory tax, which might be efficiently incorporating the negative externalities of new production, but typical estimates find that the implicit tax is far higher than most reasonable estimates of those externalities.

We begin by discussing how to estimate the minimum profitable cost of production for a house in a lightly regulated housing market, where such costs are primarily determined by geography and characteristics of local markets for labor and materials. We can then classify US housing markets into three different groups. In lightly regulated housing markets with growing population and economies, like Atlanta, the supply curve for housing is relatively flat. Thus, as demand for housing expands over time, the result is that competition in the home building industry holds the price of housing reasonably close to its minimum profitable production cost. In heavily regulated housing markets with growing economies, like the San Francisco Bay area, the supply curve for housing slopes up. As a result, additional demand for housing translates into prices that are substantially above the minimum profitable production cost, with rising land values driving up total costs. Finally, in a housing market like Detroit where the demand for housing declined sharply over time, the supply curve for housing has a kink at the existing level of housing because housing is durable and does not diminish quickly when demand falls. As a result, a reduction in demand leads to lower prices for housing and minimal new construction (Glaeser and Gyourko 2005).

The ratio of price to minimum profitable construction cost is akin to Tobin's q, the standard ratio of market value to firm replacement cost. Regulatory construction constraints can explain why this variant of q may be higher than one in some housing markets, just as capital adjustment costs can explain why q is higher than one in classical investment models (Hayashi 1982).

We then discuss two main effects of developments in housing prices: on patterns of household wealth and on the incentives for relocation to high-wage, high-productivity areas. Binding supply-side restrictions shape the personal portfolios of millions of Americans, and much of the rise in the capital share can be

[1] Policymakers often discuss housing prices through the prism of affordability: for example, many federal programs deem that housing is inappropriately expensive or unaffordable if the monetary costs of occupying your home exceed 30 percent of one's gross income. The social merits of this cutoff as a rough rule of thumb aside, economically, it lacks clarity about the extent to which the issue of housing affordability for a given area is due to a higher prevalence of households near or below the poverty level, or due to housing prices that are at relatively very high levels. It also fails to consider an implication of the standard spatial equilibrium model used in urban economics, which is that equalizing utility levels across space implies that housing costs will be a higher share of earnings in higher-wage locations (Rosen 1979; Roback 1982).

attributed to rising rents on housing. However, only a small sliver of America is sitting on a large amount of housing wealth. We will argue that the rise in housing wealth is concentrated in the major coastal markets that have high prices relative to minimum production costs, and it is concentrated among the richest members of the older cohorts—that is, on those who already owned homes several decades ago, before binding constraints on new housing construction were imposed. In effect, the changes in housing wealth reflect a redistribution from buyers to a select group of sellers.

The restrictions on housing supply and corresponding high housing prices in certain areas are also a distortion that limits the movement of workers in areas with high productivity and high wages—and also high housing costs. Hsieh and Moretti (2017) have estimated that real GDP could be nearly 9 percent higher if there were plentiful new construction in just the three high productivity markets of New York, San Francisco, and San Jose, so that people could move to equalize wages. We will discuss the basis for such estimates and show that there can be a fairly wide range of outcomes depending upon model and parameter assumptions. However, our analysis indicates that a lower bound cost of restrictive residential land use regulation is at least 2 percent of national output. If these regulatory distortions are efficiently internalizing negative externalities, then the benefit of increased aggregate output would also need to be weighed against the costs of local disamenities.

In the conclusion, we turn to some policy implications. The available evidence suggests, but does not definitively prove, that the implicit tax on development created by housing regulations is higher in many areas than any reasonable negative externalities associated with new construction. Consequently, there would appear to be welfare gains from reducing these restrictions. But in a democratic system where the rules for building and land use are largely determined by existing homeowners, development projects face a considerable disadvantage, especially since many of the potential beneficiaries of a new project do not have a place to live in the jurisdiction when possibilities for reducing regulation and expanding the supply of housing are debated.

Construction Costs and Regulations in Housing Markets

Variation across Physical Geographies in the Cost of Supplying a Home

There is no reason to expect that the production costs of housing should be the same across markets, even if those places have similar levels of regulation. Geography will make housing more expensive to build in some areas than others. Bedrock makes it easier to build up (Rosenthal and Strange 2008). Steep ground makes it much more challenging to build (Saiz 2010). Bodies of water can limit land supply.

The flat cities of the American Midwest are close to the perfect physical environment for building, as is much of the Sunbelt region. Conversely, America's coastal cities are considerably more difficult geographical environments for builders. California cities often have significant changes in elevation within a single metropolitan

area. Both East Coast and West Coast cities are limited in that they can only expand inland. All of America's oldest cities were built on major waterways because of the advantages of access to water-borne transportation. Consequently, the central business districts of markets such as Boston, New York, San Francisco, and even Chicago are close to the waterfront. Developers in those places only have a semi-circle of land to develop. The island of Manhattan poses particularly unique challenges.

When supply of housing is relatively lightly regulated, as it is throughout much of the American Sunbelt and the interior of the country, construction seems to be close to a constant-returns-to-scale technology. This relationship reflects the relative abundance of building materials such as wood, and less-skilled construction workers.[2] Of course, construction costs do vary according to the physical geography of local building conditions, but Gyourko and Saiz (2006) examine the heterogeneity of construction costs (discussed in more detail below) and find that the variance of such costs is much smaller than the heterogeneity of housing prices. This implies that we can talk sensibly about a single production cost.

Variations in Regulations on Land Use and Building

The United States is relatively unique in that land use is under local control, which leads to wide variation in regulation across communities. Many other countries, including the United Kingdom and France, have national planning agencies and guidelines set by their central governments. The type of local land use regulation in the United States, ranging from building code requirements to strict limits on the number of units delivered, also differs across markets and can affect construction costs associated with putting up the structure, as well as the underlying price of land.

Modern land use regulation in the United States dates back at least to the 1910s, when the initial zoning laws were enacted to limit negative externalities from spillovers between different kinds of land users. While there are no consistent time series measures of the local residential land use regulatory environment, researchers generally agree that such regulation has proliferated across markets and become onerous in some places. The term NIMBYism ("not in my back yard") dates back to Frieden (1979). The literature on this topic is now voluminous and Gyourko and Molloy (2015) provide a recent review.

There is no doubt that binding density restrictions affect supply. For example, the median Boston suburb has a minimum lot size over one acre—and larger minimum lot sizes are common. Unsurprisingly, minimum lot size is strongly negatively correlated with new building across communities in greater Boston (Glaeser and Ward 2009).

Restrictions often go far beyond minimum acreage or maximum height restrictions. Examples include laws that prohibit multifamily dwellings, stop development

[2]Taller buildings also display their own constant returns to scale because the per-square-foot cost of building to seven stories is quite close to the per-square-foot cost of building 50 stories (RSMeans 2015). That said, the cost of building up is much higher than the cost of building low-rise dwellings.

near wetlands (which are often loosely defined), and make it difficult to build across large swaths of historic neighborhoods. Since the 1972 *Friends of Mammoth* case, the California Environmental Quality Act (CEQA) has been interpreted to require an environmental impact review for "most proposals for physical development in California" (http://resources.ca.gov/ceqa/more/faq.html#when). Environmental impact reviews may not ultimately prevent a project, but they will add time delays that increase development costs. Moreover, the environmental impact reviews only investigate the project's impact on the local environment and do not include the environmental benefits of building in California, where carbon emissions would be low due to the mild climate, rather than, say, Texas or Arizona where they would be higher (Glaeser and Kahn 2010).

The potential for a multiyear review process, which is not uncommon in many jurisdictions, is associated with higher project uncertainty, not just time delays. A project may be denied approval after many years of active planning. That risk also increases the expected costs to developers and deters new housing supply.

The plethora of restrictions on building makes it difficult to measure the overall strictness of the broader regulatory environment, but it is possible to describe the nature of different types of communities' approaches to regulation. The Wharton Residential Land Use Regulation Index, based on surveys of local government officials, documents wide differences in the difficulty of obtaining building permits across metropolitan areas (Gyourko, Saiz, and Summers 2008). The typical regulatory environment in their sample of 2,611 communities across 293 metropolitan areas can be described as follows: 1) two entities are required to approve any project requiring a zoning change, so there are multiple opportunities for rejection; 2) minimum lot size restrictions are omnipresent; 3) "development exaction fee programs" also are now omnipresent; and 4) the typical community exhibits about a six-month lag between submission of a permit request for a standard project and a decision on whether to approve it.

The one-third most highly regulated communities in the Gyourko, Saiz, and Summers (2008) sample also share some additional traits. Local and state pressure groups are much more likely to be involved in the regulatory process in these communities. More than half the highly regulated places have at least one neighborhood with a one-acre (or more) minimum lot size rule; in contrast, only 5 percent of the one-third most lightly regulated communities had any neighborhood with a one-acre minimum rule. Open space requirements, not just development exactions, are now common in highly regulated places. Finally, the most highly regulated places have project approval lags that average ten months in length, which is three times longer than in the least regulated one-third of communities. In another study, Saiz (2010) documents how both regulations and geography limit building and increase prices across space.

A variety of models of local land use control embed the idea that not all local residents will share the same goals, so that the regulatory environment will be shaped by the incentives and influence of different actors in the political process. For example, Fischel (2001) emphasizes the role of existing homeowners, who have

a strong incentive to protect what often is their most important asset. One obvious way to protect asset value is to restrict new supply. Theoretical analysis is much more challenging in a multicommunity setting that permits Tiebout-style sorting and strategic interactions (for discussion, see Gyourko and Molloy 2015). In principle, regulation can be an efficient means of forcing developers to internalize negative externalities from construction. Moreover, the spatial heterogeneity in those regulations may reflect different external costs from construction, perhaps because of different local preferences.

The general conclusion of existing research is that local land use regulation reduces the elasticity of housing supply, and that this results in a smaller stock of housing, higher house prices, greater volatility of house prices, and less volatility of new construction. Most results are consistent with these implications, and we report additional evidence below. However, it has been a challenge in this literature to find convincing instruments or some form of experimental variation. Because empirical work in this area is cross sectional in nature, it is subject to standard potential biases associated with omitted variables and reverse causality.[3]

What Does It Actually Cost to Supply Homes to the Market?

There are three components to the cost of delivering a unit of housing to the market: 1) the land (L) on which the housing unit sits; 2) construction costs (CC) associated with putting up structure itself; and 3) a rate of entrepreneurial profit (EP) needed to compensate the home builder. Thus, we define the "minimum profitable production cost" (MPPC) of a unit of housing as follows:

$$\text{MPPC} = (\text{L} + \text{CC}) \times \text{EP}.$$

Vacant land sales are rarely observed in the United States, so to estimate the value of a price of land, we use an industry rule of thumb based on an *ad hoc* survey of home builders that land values are no more than 20 percent of the sum of physical construction costs plus land in a relatively free market with few restrictions on building. We have used this metric in earlier research (Glaeser and Gyourko 2003, 2008), and it continues to be relevant and consistent with the data discussed below.

The gross profit margin on the builder's land and construction costs for a portfolio of homebuilders range from 9–11 percent per annum across the cycle. This implies gross margins of about 17 percent given that cost of operations are roughly 35–40 percent of gross margins for such companies. Hence, EP = 1.17 in our calculations below.

[3]To understand the problem of finding experimental variation in this literature, consider the variation in difficulty of building across space generated by geographic variables of the type analyzed by Saiz (2010). In this setting, a location that is close to water increases housing demand, but also creates a more challenging geographical environment. More generally, home-building will occur in more challenging and costly locations only if those locations have something else going for them. Consequently, geography provides meaningful variation in the difficulty of building, but is not a valid instrument for housing supply in most situations.

Physical construction costs are more readily observable from the home building industry. We use RS Means Company (RSMeans) data on physical construction costs as the foundation of our estimates of minimum profitable production cost. This firm provides and sells estimates of the cost of providing units of different qualities across more than 100 American housing markets. Their data have been used by us (and others) in previous research (for examples, see Glaeser and Gyourko 2003, 2005; Glaeser, Gyourko, and Saks 2005; Gyourko and Saiz 2006).

The RSMeans cost estimates cover material, labor, and equipment (but not land) for four different qualities of single family homes—economy, average, custom, and luxury. Means reports costs per square foot and provides estimates for homes ranging in size from 600 ft^2 to 3,200 ft^2 of living area. Breakdowns are available by the number of stories in the house, and certain other characteristics (such as the presence of a basement). We focus on costs associated with a smaller, modest-quality, one-story home of economy quality described in RSMeans publication, *Residential Cost Data 2015*.[4] We choose this home because we believe it reflects the quality of the typical home (which is not new or very large) in most, if not all, markets. We have experimented with using this data with regard to homes of other quality characteristics and discuss possible biases below.

The first important fact is that structure costs are modest for an economy-quality home. The interquartile range runs from \$72/ft^2 to \$86/ft^2, and the distribution is not fat-tailed. The 5^{th} and 95^{th} percentile values are \$68/$ft^2$and \$95/ft^2, respectively. Thus, in cheaper markets, physical construction costs associated with putting up a typical home with 2,000 ft^2 of living space are about \$140,000 (approximately \$70 per square foot); in the most expensive markets, the costs are about \$180,000 (approximately \$90 per square foot).

A second noteworthy fact is that real construction costs have not risen much over time. Measured in constant 2010 dollars, the cost was \$83 per square foot in 1980, had declined slightly to the mid-\$60s per square foot by the late 1990s and early 2000s, and then rose back to \$85 per square foot by 2015. This finding is consistent with much previous research and implies that rising real house prices cannot be explained by higher physical construction costs (for example, Davis and Heathcote 2007; Davis and Palumbo 2008; Gyourko and Molloy 2015).

These relatively constant physical production costs help us to understand the often-noted decline in total factor productivity in the construction sector (for example, Barbosa et al. 2017). This decline does not seem to result from any change in building technology, but rather an increase in other costs associated with delivering housing, such as dealing with regulation.

[4]Specifically, this is a one-story single family home, one full bathroom, one kitchen, asphalt roof shingles, hot air heat, gypsum wallboard interior finishes, mass produced from stock plans. The RSMeans Company presumes that a given quality home is constructed in a common way across markets. It divides the home into a number of different tasks that require certain services, materials, or labor. RSMeans then surveys local suppliers and builders to determine the local price of those inputs. One-off construction of custom homes would be much more costly. See RSMeans Company (2015) and section 2 in Gyourko and Saiz (2006) for more on the underlying methodology.

Given the assumptions outlined above for costs of land and profits, minimum profitable production costs that take land and profit into account are nearly 50 percent higher than the RSMeans physical construction cost numbers. This suggests that an efficient housing market should be able to supply economy-quality single-family housing with 2,000 ft^2 of living space for around $200,000 in low construction cost markets and for little more than $265,000 in the highest construction cost markets. The key factors that account for the cross-sectional variation in structure production costs in this data are the extent of unionization in the construction industry, the level of local wages in general, and difficult topography (Gyourko and Saiz 2006). For perspective, what RSMeans calls the "average" quality home costs about 25 percent more than the economy home, and the highest quality "luxury" home of the same size costs almost twice as much to construct as the economy home.

Comparing Minimum Profitable Production Cost and Actual Housing Prices

We can compute the ratios of house prices to the minimum profitable production cost using different data sources on home values. Most of our results below use self-reported house prices from the microdata in the biannual American Housing Survey (AHS), which runs from 1985 to 2013. It reports data on individual housing units and their occupants in 98 core-based statistical areas (CBSAs), which refers to a metropolitan area of one or more counties anchored by an urban center of at least 10,000 people, tied together by commuting patterns. These markets (which are listed in an online Appendix to this article at http://e-jep.org) contain approximately 75 percent of the urbanized population in the United States according to 2010 Census data and include virtually any market of significant size.[5]

Some strengths of the American Housing Survey data are that they contain microdata, clearly identify single-family detached units, and report the square footage of living area. The latter is useful as it allows us to match units of different sizes with the appropriate construction cost in the RSMeans data. (Smaller units typically have higher costs per square foot.) We do this for homes of 600, 800, 1,000, 1,200, 1,400, 1,600, 1,800, 2,000, 2,400, 2,800, 3,200, 3,600 and 4,000+ square feet of living area. Specifically, if a house is reported to be less than or equal to 700 square feet of living area, this is matched to RSMeans costs per square foot for a 600 square-foot, economy-quality home.

Each single family home that includes data on living area is matched with cost data from RSMeans and then grouped into one of four bins, based on the ratio of housing prices to minimum profitable production cost (P/MPPC): 1) A ratio of 0.75 or less, which implies that market value of the house is at least 25 percent below

[5]We cannot calculate a truly national ratio of housing price-to-minimum profitable production cost. Construction costs are not reported by RSMeans for each market in the country, and no such data are available for rural areas either. Moreover, the American Community Survey does not report anything on housing unit size, which means that added assumptions need to be made if using its data to compare housing prices and costs. We did experiment with the median-priced-unit from the 2014 American Community Survey in computing price-to-cost ratios like those we are about to discuss. Our findings are very similar in quality and quantity to those we will report using the American Housing Survey.

our estimate of reproduction costs; 2) a ratio between 0.75 and 1.25, which we interpret to be the range within which prices are not materially different from minimum profitable production costs; 3) a ratio between 1.25 and 2; and 4) a ratio greater than 2, which implies that prices are more than double our estimate of production costs. We chose these four relatively wide bins because they are likely to be reasonably robust to the measurement error involved in the construction of our ratios.

These ratios are essentially the value of Tobin's q for housing. Just as in standard investment theory, a value of q below one implies that the capital would not be replaced if it were destroyed. Values of q above one must reflect some barrier to investment, which we believe is more likely to be regulation in the housing market rather than standard adjustment costs (Hayashi 1982). Values of q above one can also be a sign of market overvaluation, as in Las Vegas in 2005, but only in cases where land is abundant and regulations are few.

Table 1 reports our baseline results, which include data from 1985 to 2013. As of 2013, slightly less than three-quarters of all observations (73.6 percent) are priced near or below minimum profitable production costs (we see this by adding together numbers from the first two columns), with more than half of them being valued more than 25 percent below. This leaves just over one-fourth (26.4 percent) living in expensive housing, with 10 percent of the underlying sample living in homes estimated to be more than double minimum profitable production costs. In a large swath of urban America—and especially if one focuses on the local housing markets in the bottom four-fifths of prices—the housing market is supplying units at quite reasonable prices given all-in production costs.

Also, Table 1 shows that the housing cycle matters. For example, at the height of the last housing boom, the 2005 data indicate that more than one-half of all observations were at least 25 percent more expensive than minimum profitable production costs.[6]

Given the inevitable measurement error arising from unobserved quality differences across housing units reported in the microdata, another way to examine the spatial distribution of housing prices is at the metropolitan-area level. We look at the ratio of the median housing price to the minimum profitable production cost in every housing market for which we have at least 25 individual observations.[7] Table 2

[6]We also experimented with different housing quality assumptions in computing minimum profitable production cost. Using the lowest quality that meets local building codes for the cost of supplying a home will result in misclassifying some observations as expensive (that is, with a ratio over 125 percent), especially those living in elite suburbs. If we use the costs associated with what RSMeans terms "average" quality (one above economy quality), the share of observations classified as expensive falls from 26 to 18 percent. Using the highest possible construction quality—the "luxury" homes in RSMeans terminology—is required to dramatically lower the estimate of expensive homes. In that case, the share of observations valued at more than 125 percent of minimum profitable production cost falls to just over 6 percent. Thus, our conclusion that the vast majority of homes are priced near or below their full social costs of replication is robust to virtually any assumption we could make.

[7]This use of the median only for markets with 25 or more observations results in an unbalanced panel of markets, but the findings are not materially different if we restrict the data to the common set of metropolitan areas for which we have at least 25 observations each survey year dating back to 1985.

Table 1
House Price to Minimum Profitable Production Cost Ratio (P/MPPC): Using All the Micro Data
(percent of observations that fall into each category)

Year	$P/MPPC \leq 0.75$	$0.75 < P/MPPC \leq 1.25$	$1.25 < P/MPPC \leq 2$	$P/MPPC > 2$
1985	38.0%	40.5%	17.9%	3.6%
1987	33.4%	38.3%	21.7%	6.6%
1989	31.8%	34.6%	20.3%	13.3%
1991	31.1%	35.3%	22.5%	11.1%
1993	31.8%	36.1%	23.6%	8.5%
1995	27.4%	37.7%	26.5%	8.4%
1997	31.5%	40.0%	23.0%	5.5%
1999	22.0%	40.1%	26.2%	11.8%
2001	19.4%	38.2%	25.2%	17.1%
2003	16.2%	32.1%	25.9%	25.9%
2005	18.0%	28.7%	25.3%	28.0%
2007	19.9%	28.1%	24.0%	28.0%
2009	31.4 %	33.9%	21.6%	13.1%
2011	37.4%	35.4%	16.0%	11.2%
2013	40.3%	33.3%	16.2%	10.2%

Source: The calculations are based on self-reported house prices from the micro data in the biannual American Housing Survey (AHS), which runs from 1985 to 2013 and reports data on individual housing units and their occupants in 98 core-based statistical areas (CBSAs), containing approximately 75 percent of the urbanized population in the United States. Each single family home from the AHS that includes data on living area is matched with data on construction costs from RSMeans.
Notes: The table shows the percentage of single family homes each year that fall into one of four bins based on the ratio of housing prices to minimum profitable production cost (P/MPPC): 1) A ratio of 0.75 or less, which implies that market value of the house is at least 25 percent below our estimate of reproduction costs; 2) a ratio between 0.75 and 1.25, the range within which prices are not materially different from minimum profitable production costs; 3) a ratio between 1.25 and 2; and 4) a ratio greater than 2, which implies that prices are more than double our estimate of production costs.

shows the percentage of metropolitan areas each year that fall into one of four bins based on the ratio of housing prices to minimum profitable production cost (P/MPPC).

In 1985, over 90 percent of our metropolitan areas had median price-to-cost ratios less than or near 1. Only five (6.4 percent) had medians above 1.25 (and there were none where price was more than double production cost). The percentage of observations that year in the two most-overpriced categories on Table 2 is only one-third of the 21.5 percent reported for these same two categories in Table 1, which uses all the microdata. But we do not find this surprising given the measurement error issue associated with unobserved unit quality, especially for homes located in the most desired suburbs. As of the middle of the 1980s, in only a handful of markets concentrated in California and Hawaii, and none on the East Coast, was the typical home expensive relative to minimum profitable production cost. Based on earlier Census data, we presume that this distribution largely characterized housing markets before that point as well (Gyourko, Mayer, and Sinai 2013).

Table 2

House Price-to-Minimum Profitable Production Cost Ratio (P/MPPC): Median Values across Core-Based Statistical Areas (CBSAs)

(percent of observations that fall into each category)

Year	*Number of CBSAs*	*P/MPPC ≤ 0.75*	*0.75<P/MPPC ≤ 1.25*	*1.25<P/MPPC ≤ 2*	*P/MPPC > 2*
1985	78	37.2%	56.4%	6.4%	0.0%
1987	73	28.8%	57.5%	13.7%	0.0%
1989	78	34.6%	50.0%	10.3%	5.1%
1991	71	23.9%	57.7%	14.1%	4.2%
1993	79	25.3%	62.0%	11.4%	1.3%
1995	72	19.4%	68.1%	9.7%	2.8%
1997	70	15.7%	71.4%	12.9%	0.0%
1999	71	8.5%	74.6%	14.1%	2.8%
2001	71	7.0%	69.0%	16.9%	7.0%
2003	71	5.6%	60.6%	23.9%	9.9%
2005	70	11.4%	44.3%	27.1%	17.1%
2007	66	12.1%	39.4%	30.3%	18.2%
2009	65	24.6%	50.8%	20.0%	4.6%
2011	67	29.8%	50.7%	14.9%	4.5%
2013	69	33.3%	50.7%	10.1%	5.8%

Source: The calculations are based on self-reported house prices from the microdata in the biannual American Housing Survey (AHS), which runs from 1985 to 2013 and reports data on individual housing units and their occupants in 98 core-based statistical areas (CBSAs), containing approximately 75 percent of the urbanized population in the United States. Each single family home from the AHS that includes data on living area is matched with data on construction costs from R.S. Means Company.

Notes: We look at the ratio of the median housing price to the minimum profitable production cost in every housing market (core-based statistical area) for which we have at least 25 individual observations. The table shows the percentage of metropolitan areas each year that fall into one of four bins based on the ratio of housing prices to minimum profitable production cost (P/MPPC): 1) A ratio of 0.75 or less, which implies that market value of the house is at least 25 percent below our estimate of reproduction costs; 2) a ratio between 0.75 and 1.25, the range within which prices are not materially different from minimum profitable production costs; 3) a ratio between 1.25 and 2; and 4) a ratio greater than 2, which implies that prices are more than double our estimate of production costs.

During the late 1980s boom in housing prices, median prices shifted up relative to construction costs. By 1991, the share of metropolitan areas with median value-to-cost ratios below 0.75 had fallen to 24 percent, but another 58 percent had values reasonably close to 1. The share of metropolitan areas with median price-to-cost ratios greater than 1.25 more than doubled to just over 14 percent, with the Honolulu, Los Angeles, and San Francisco markets having prices more than double minimum profitable production costs.

The mid-1990s seems to have been a time of compression of metropolitan area prices, just as it was the only period in recent decades in which income inequality also declined. But between 1997 and 2007, median price-to-cost ratios in the most expensive markets rose dramatically. At the height of the boom in 2007, just over 48 percent of our metropolitan areas had median ratios with prices more than 25 percent above estimated reproduction costs, with one-third of those areas having price-to-cost ratios that were greater than two.

The years following the global financial crisis saw a distribution of median price-to-cost ratios that looked much like the early 1990s. By 2013, only three markets had median price-to-cost ratios above 2—the same number as in 1991. Nearly 10 percent had ratios between 125 and 200 percent, which is only slightly lower than the analogous share in 1991. Median price-to-cost ratios were less than 0.75 in one-third of markets in 2013, which is higher than the 24 percent in 1991. This pattern implies that in a substantial fraction of urbanized America, it would not pay to rebuild the typical home if it fell down today. Nominal prices have gone up in these areas since the late 1980s, but nominal construction costs have risen as well.

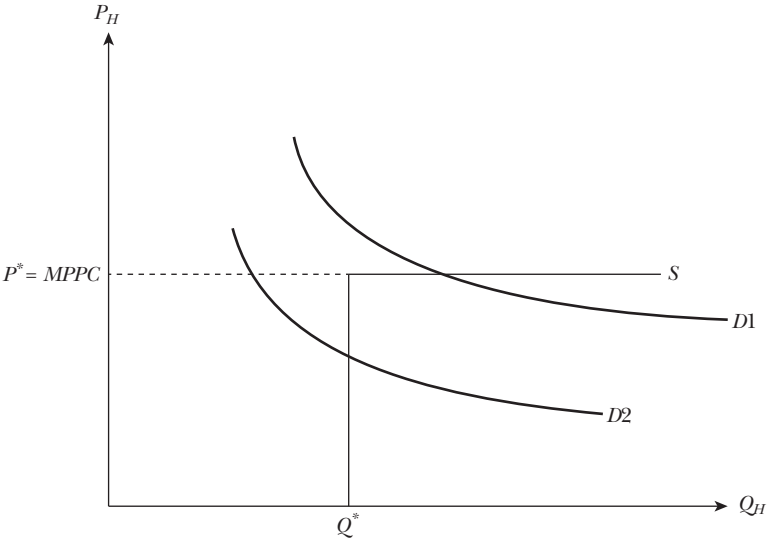
Perhaps the largest difference between 1985 and 2013 is that the share of metropolitan areas with median price-to-cost ratios above 1.25 has risen from 6.4 to 15.9 percent. There are now a modest number of markets in America (though more than in 1985) where the typical owner is living in a home that is priced substantially above minimum profitable production costs. These markets include some of the nation's most productive labor markets, so they are important for the nation's economic future.[8]

This gap between price and cost seems to reflect the influence of regulation, not the scarcity value arising from a purely physical or geographic limitation on the supply of land. For example, in Glaeser, Gyourko, and Saks (2005), we show that the cost of Manhattan apartments are far higher than marginal construction costs, and more apartments could readily be delivered by building up without using more land. This and other research we have done (Glaeser and Gyourko 2003) also finds that land is worth far more when it sits under a new home than when it extends the lot of an existing home, which is also most compatible with a view that the limitation is related to permits, not acreage per se.

It is possible that regulatory limits on construction are efficiently internalizing the negative externalities from construction, but the vast gap between price and construction cost in some coastal markets could only be justified by enormous construction externalities. Empirical investigations of the local costs and benefits of restricting building generally conclude that the negative externalities are not nearly large enough to justify the costs of regulation (Cheshire and Sheppard 2002; Glaeser, Gyourko, and Saks 2005; Turner, Haughwout, and van der Klaauw 2014). Glaeser and Ward (2009) also find that the impact of neighborhood-level density on housing values in greater Boston is far too small to justify the current restrictions on new construction.

[8] The three markets with ratios of median housing price to minimum profitable production cost above 2 are Los Angeles–Long Beach–Anaheim, CA; Oxnard–Thousand Oaks–Ventura, CA; and San Francisco–Oakland–Hayward, CA. Those with ratios between 1.25 and 2 are Baltimore–Columbia–Towson, MD; Boston–Cambridge–Newton, MA–NH; Denver–Aurora–Lakewood, CO; New York–Newark–Jersey City, NY–NJ–PA; San Diego–Carlsbad, CA; Seattle–Tacoma–Bellevue, WA; and Washington–Arlington–Alexandria; DC–VA–MD–WV.

Figure 1
Kinked Supply Schedule from Durable Housing



Note: "MPPC" means minimum profitable production costs.

A Closer Look at Three Types of Markets: Detroit, Atlanta, and San Francisco

The housing market in Detroit–Warren–Dearborn, Michigan, is emblematic of a place in which home prices have been well under minimum profitable production costs for long periods of time. This is one of the cases illustrated in Figure 1. The supply schedule of housing is always kinked, with the vertical component reflecting the size of the current stock. The height of the supply schedule at the kink is minimum profitable production costs. Even as the housing market in Detroit–Warren–Dearborn was growing in the past, prices were pinned down by minimum profitable production cost (Glaeser and Gyourko 2005), as shown by the intersection of supply and demand, D_1, which is on the horizontal part of the supply schedule.

Following a negative demand shock for the market (in this case, fierce foreign competition for the domestic auto industry, which was concentrated in Detroit), demand dropped to D_2 and now intersects supply on its vertical component. Prices are below the full production cost of new housing, because this intersection reflects the depreciated price of older housing. Most Americans, not just those in declining markets, do not live in new units. More than seven million occupied housing units were built before 1919, constituting approximately 6.2 percent of the occupied housing stock. Over 30 percent of occupied units in 2014 were built before 1960 and so were more than 50 years old. As shown in Figure 2A, the ratio of house prices to minimum profitable production cost in Detroit was well below 1 for much of the 1980s and 1990s, and then rose towards 1 during the recent long boom, before falling back after the bust ensued. Unsurprisingly, annual building permits in Detroit are

not more than about 1 percent of the market's 2000 housing stock in any year since 1985—and were near zero from 2007–2011, according to American Housing Survey data.

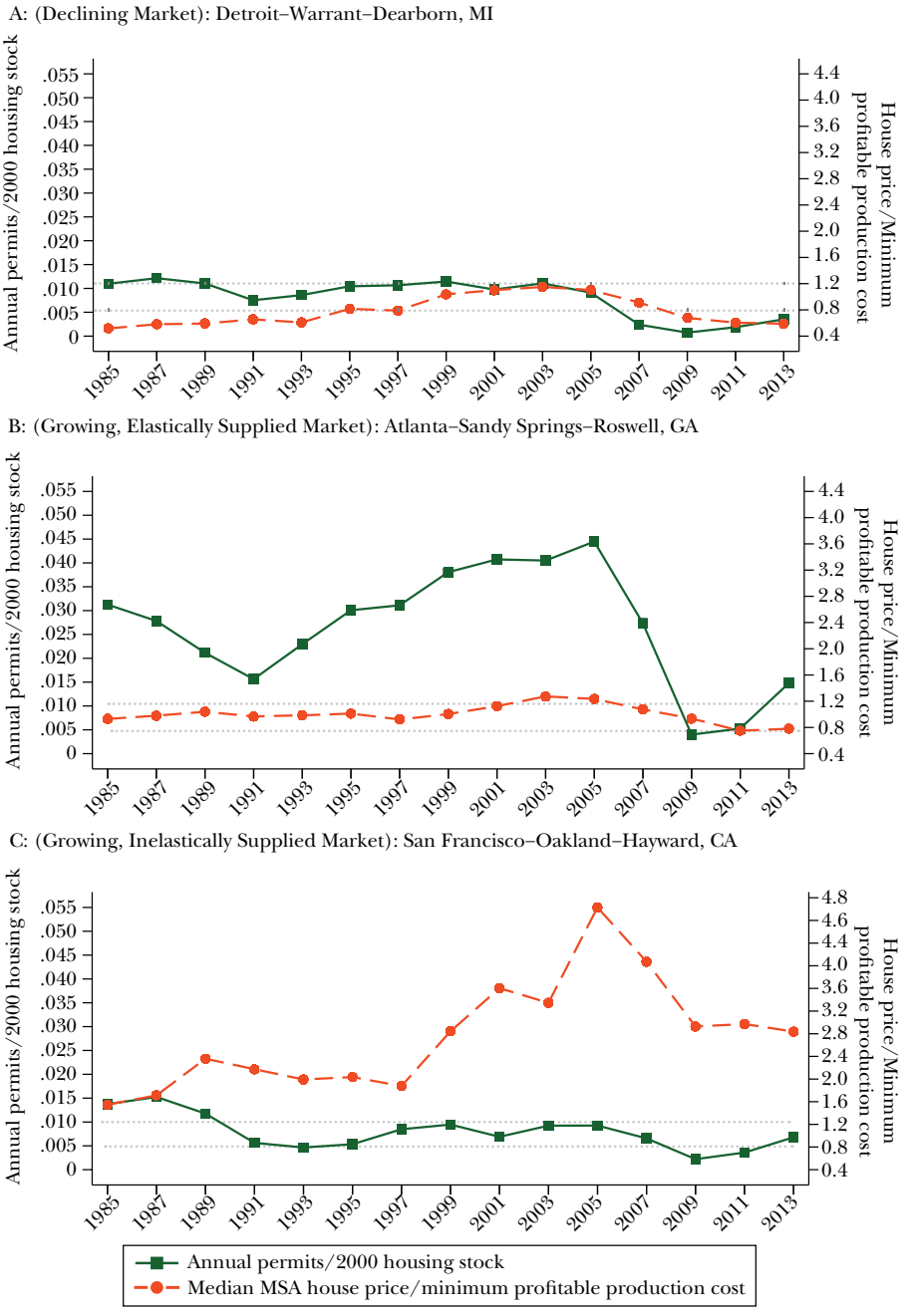
The Atlanta market is a canonical example of a local housing market in which supply is highly elastic (beyond the kink) and demand is strong enough to always intersect the supply schedule beyond the kink, keeping prices at minimal profitable production cost. As Figure 2B shows, new supply is highly volatile. Permitting intensity was running at 3 percent of market size in 1985; fell half by 1991 as the local economy declined; more than doubled to nearly 4.5 percent of market size by 2005; plummeted to below 0.5 percent of market size in the throes of the financial crisis by 2009; and has only recently started to increase again. Amidst all this variation in new supply, the median owner's price-to-cost ratio never varies much from 1. This is consistent with a highly elastic supply side of the housing market, in which demand is intersecting supply beyond the kink on the horizontal part of the supply schedule in Figure 1.

San Francisco represents the third type of housing market in which the price of housing is considerably above the minimum profitable production cost. In this situation, strict regulation of housing construction means developers in this type of market cannot bring on new supply even though it looks as if they could earn super-normal profits if they did. Unlike the graph in Figure 1, the supply schedule beyond the kink is upward sloping, and demand is strong enough to intersect the supply schedule well above where P = MPPC. Thus, shifts in the demand for housing affect price more than quantity. As Figure 2C shows, the median house price in this market has been well above the minimum profitable production cost for the past three decades and reached dramatic heights at the peak of the last housing boom in 2005. However, permitting activity did not increase at all over the eight-year span from 1997 to 2005, even though the median price-to-cost ratio increased from below 2 to over 5. Although the ratio has fallen sharply from that peak, it remained a very high 2.84 as of 2013. The link between prices (relative to production costs) and new supply has been broken in this type of market.

San Francisco is a relatively high physical construction cost market, but that is not what makes its homes cost so much. The median housing unit in this market contained 1900 square feet, and the physical construction costs for this unit based on RSMeans data were $192,938, so the per square foot cost of the (presumed modest quality) structure was just over $100 per square foot, which is one of the most expensive construction cost markets in the United States. Our earlier assumption that land is 20 percent of the physical-cost-plus-land total provides an estimated land price of $48,235. Stated differently, that is what we think the underlying land would cost in a relatively unregulated residential development market. Add the builder's 17 percent gross margin, and the minimum profitable production cost for this house is $281,690. This compares with an actual price of the median house of $800,000 (and thus a price-to-cost ratio of 2.84).

Clearly, San Francisco housing developers cannot actually earn super-normal profits on the margin. Instead, what makes San Francisco housing so expensive is

Figure 2
New Housing Supply and House Prices (Relative to Costs)



Source: House prices come from American Housing Survey micro data.
Note: Minimum Profitable Production Costs (MPPC) are for 1800 square feet, economy class, 20 percent land share, and 17 percent gross margin homes.

the bidding up of land values. Our formula suggests that the land underlying this particular modest-quality home cost about $490,000—roughly 10 times the amount presumed for our underlying calculations of the minimum profitable production cost.

The time path of prices in the three cities is representative of a larger pattern: cities with inelastic housing supply generally experienced much more extreme price gyrations during the boom–bust cycles of the 1980s and 2000s. In Glaeser, Gyourko, and Saiz (2008), we report that in the 1980s boom, mean price growth was 29 percent for most inelastic metropolitan areas and 3.4 percent for the most elastic metropolitan areas. During the 1996–2006 boom, mean real price growth was 93.9 percent in the most inelastic cities and 28.2 percent in the most elastic cities. The remarkable element in the 1996 to 2006 period is that some relatively elastic cities, such as Phoenix and Las Vegas, still experienced extremely high price growth over a short time period, and equally sharp subsequent declines.

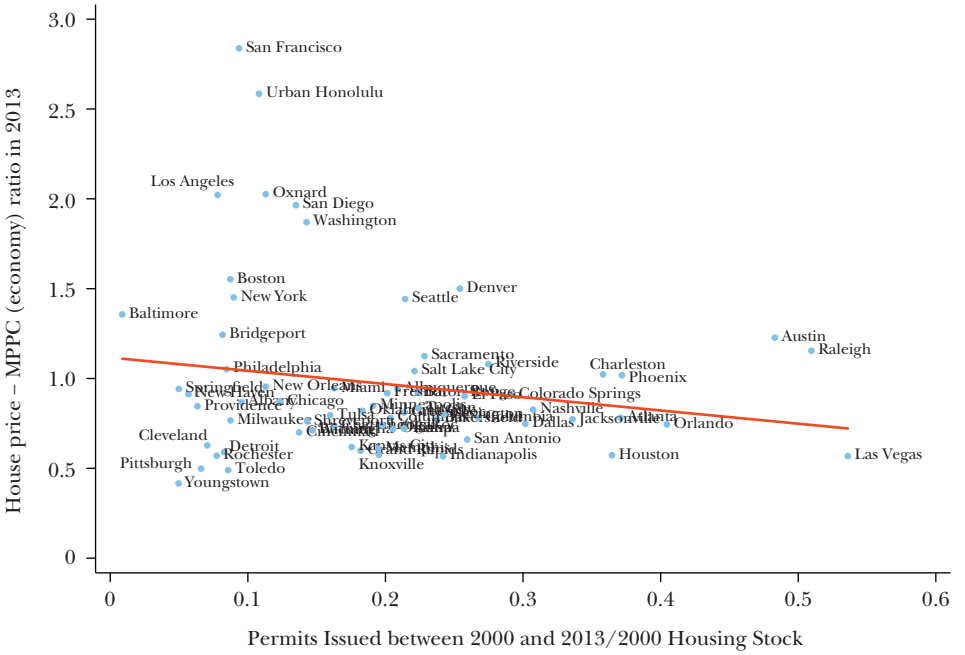
Additional Connections

Overall, more expensive housing markets tend to be both more regulated and have more inelastic supply sides. The correlation of median house price in 2013 with the Wharton Residential Land Use Restrictiveness Index (which has a bigger value the more restrictive the regulatory environment) is about 0.5. This is very similar to the magnitude of the correlation with Saiz's (2010) elasticity measure (although of the opposite sign because his measure declines in value when supply is more inelastic).

A broader look at our data also shows a clear connection between housing prices and new home construction activity. Figure 3 confirms that Atlanta, Detroit, and San Francisco are, indeed, representative of the three market types discussed. Price-to-MPPC ratios in 2013 are plotted against the magnitude of construction activity as reflected by the ratio of new units built between 2000 and 2013 to the housing stock in 2000. The modest negative slope that best fits that scatter plot of markets is driven by the following combination of facts: 1) among markets with high P/MPPC ratios of 1.5+, there was relatively little new home construction over this 13-year period (typically less than 15 percent in aggregate, or about 1 percent per annum on a compounded basis), and in addition, there is little variation in permitting intensity among this group of the most expensive housing markets; 2) among markets with low P/MPPC ratios of 0.7 or less, there also was very little new home construction, so that building intensity in the lowest-price housing markets (Detroit, Cleveland and Rochester) is not much less than in Boston or New York City—although the reason is that builders cannot earn a profit in the low-price markets; and 3) among the markets with P/MPPC ratios closer to one, there is a much wider range of building levels, depending upon the level of demand in each metropolitan area.

Our data also shows a marked increase in price dispersion across markets, with the right tail of inflation-adjusted housing prices much longer now than it was three decades earlier. This is consistent with earlier research (for example, Gyourko, Mayer, and Sinai 2013).

Figure 3
Price-to-Cost Ratios and Permitting Intensity, 2000–2013



Source: Housing stock data are from the 2000 decennial census. Housing permit data were collected from the Department of Housing and Urban Development's State of the Cities Data System (SOCDS) at http://www.huduser.gov/portal/datasets/socda/html. Price-to-cost ratios were created using the data discussed above in Tables 1 and 2.
Note: MPPC is Minimum Profitable Production Cost.

Finally, there is a strong correlation between homeowner income and the degree of regulation in a market. Variation in the Wharton regulatory index or Saiz's (2010) elasticity can account for nearly 25 percent of the variation in the income of the owner of the median-priced home in 2014 based on American Community Survey data. Given the aforementioned positive correlation between house prices and the degree of regulatory constraint, it is not surprising to find higher-income people living in more expensive homes. Of course, no causal relation is implied from a simple bivariate correlation. However, this does link to one of the most important new implications of inelastic supply sides in coastal markets—the potential impact on the distribution of wealth and on the geographic distribution of where people of different income levels are more likely to end up living. We now turn to these issues.

The Impact of Supply Restrictions: Household Wealth

If housing restrictions have helped cause the secular rise in coastal housing prices, and the enormous volatility of prices during boom–bust cycles, then they may help explain the movement in household wealth in the United States (and elsewhere). Piketty (2014) estimates that the ratio of the US capital stock to GDP increased from 332 percent in 1970 to 410 percent in 2010, and that increases in the value of the housing stock accounts for 40 percent of this increase. Increases in housing capital account for 83 percent of the increase in the ratio of private capital-to-income between 1970 and 2010. As Rognlie (2015) has carefully documented, the net capital share increase in the post–World War II era due to housing was from 3 to 8 percent of domestic value added. La Cava (2016) argues that this increase in housing wealth in recent decades has largely been due to supply-constrained markets.

This growth in the stock of housing capital relative to GDP in recent decades is primarily about prices, not the physical supply of housing. Between 1973 and 2010, the average new home expanded from 1,660 square feet to 2,392 square feet, but this 44 percent increase is far less than the 100 percent increase in income over the same time period. Standard indices such as the S&P/Case–Shiller Index or the Federal Housing Finance Agency (FHFA) housing price index, which use repeat sales and other methods to control for changes in the quality of housing quality, still show impressive increases in prices in restricted markets, such as the 109 percent increase in real prices in greater San Francisco between 1991 and 2016. Owners of even modest properties in San Francisco who were fortunate enough to have bought prior to the rise of restrictive building regulations have seen an increase in wealth of several hundred thousand dollars. This increase in wealth is due to higher costs of land, not higher costs of physical construction, and in turn, we believe that the higher cost of land has been driven by binding land use restrictions.[9]

Yet housing wealth is different from other forms of wealth because rising prices both increase the financial value of an asset and the cost of living. An infinitively lived homeowner who has no intention of moving and is not credit-constrained would be no better off if her home doubled in value and no worse off if her home value declined. The asset value increase exactly offsets the rising cost of living (Sinai and Souleles 2005). This logic explains why home-rich New Yorkers or Parisians may not feel privileged: if they want to continue living in their homes, sky-high housing values do them little good.

Ultimately, the source of high housing costs determines its impact on well-being and personal finances. For example, if higher housing prices reflect higher wages, then San Francisco may have become less affordable, but residents who have owned property for a time are also richer. This logic leads Moretti (2013) to conclude that nominal wage inequality overstates true inequality, because those with high incomes need to pay more for access to their well-paid labor markets. Conversely,

[9] See footnote 27 in our working paper version (Glaeser and Gyourko 2017) for the calculations behind this conclusion.

Diamond (2016) argues that high housing prices in educated metropolitan areas reflect higher amenity values in those areas, which implies that real inequality is higher than earnings inequality. More generally, if higher housing prices reflect more amenities, then buyers are no worse off, but if they reflect a greater demand for the same amenities, then buyers' welfare has fallen.

In any event, the gains in housing wealth are not evenly distributed. When housing prices rise, those who already own housing are essentially hedged against a higher cost of housing (Sinai and Souleles 2005). Renters, conversely, experience the rising housing costs directly and become poorer in real terms.

Because homeowners tend to be older while renters are younger, the limited growth in housing supply has created an intergenerational transfer to currently older people who happened to have owned in the relatively small number of coastal markets that have seen land values increase substantially. On a per-owner basis, the value of these wealth gains can be considerable, but the number of markets is relatively small and many are not particularly populous. Only 11 of our Core Based Statistical Areas have a housing price-to-cost ratio above 1.25 in 2013. In total, they contained 58.8 million people and 22.9 million total housing units (according to American Community Survey data for 2014). More than half of this total for these markets consists of the 31 million people and 12 million housing units in the huge New York City and Los Angeles markets; in total, these areas contain only about 23 percent of total urban population. This relatively low share of the urban population should not be a surprise: after all, these are areas with strong constraints on building, and people cannot move to these cities without a place to live.

Table 3 presents data on net worth for six different pairs of age groups in 1983 and in 2013 from the Survey of Consumer Finances carried out by the Federal Reserve. The public use samples do not provide any geographical identifiers, but we focus here on facts about home equity. We report values for the 50th, 75th, 90th, 95th, and 99th percentiles of the distribution. Given the aggregate sample size, there are 30–40 observations per percentile. We report the average of those observations.

This table allows us to look at the same age cohort at two different times, three decades apart: for example, comparing housing wealth for 18–24 year-olds in 1983 and in 2013. The 18–24 age group has little housing wealth in 1983, and less at each percentile level in 2013. For the intermediate age groups—25–34, 35–44, 45–54—housing wealth is lower in 2013 than in 1983 at the 50th and 75th percentiles, and either roughly the same or lower at the 90th percentile. However, housing wealth is somewhat higher for these groups at the 95th and 99th percentile in 2013. For the oldest age groups—55–64 and 65–74—housing wealth is up considerably at the 90th percentile and above, with the increases being especially notable in the oldest group. Many in these age groups established themselves as homeowners 30–40 years earlier, and so were in the best position to benefit from a rise in housing prices. In short, the Survey of Consumer Finances shows sharp home wealth increases only among the richest members of the oldest cohorts. Given the potential magnitudes involved and the rising prices in many coastal markets since the latest data from 2013, these patterns seem likely to have continued.

Table 3
Housing Net Worth—30 Year Changes
(in 2013 dollars)

	1983		*2013*	
Percentile	*18–24 year-olds*	*45–54 year-olds*	*18–24 year-olds*	*45–54 year-olds*
50	$0	$87,120	$0	$30,000
75	$0	$152,159	$0	$109,000
90	$24,803	$248,818	$5,500	$250,000
95	$47,488	$353,190	$43,000	$400,000
99	$141,808	$862,359	$95,000	$1,000,000
Percentile	*25–34 year-olds*	*55–64 year-olds*	*25–34 year-olds*	*55–64 year-olds*
50	$0	$94,184	$0	$60,000
75	$45,352	$161,886	$21,000	$167,000
90	$91,827	$255,361	$74,000	$350,000
95	$123,135	$353,190	$140,000	$543,000
99	$230,751	$760,380	$256,000	$1,500,000
Percentile	*35–44 year-olds*	*65–74 year-olds*	*35–44 year-olds*	*65–74 year-olds*
50	$55,799	$82,411	$6,000	$100,000
75	$118,660	$150,136	$58,200	$225,000
90	$180,763	$279,972	$168,000	$440,000
95	$247,349	$426,936	$300,000	$701,000
99	$531,198	$941,840	$1,025,000	$2,000,000

Notes: Data compiled from the 1983 and 2013 Survey of Consumer Finances using publicly available samples.

The big winners from the reduction in housing supply are a small number of older Americans who bought when prices were much lower. Some of this wealth may be passed to the next generation as bequests. But much of the housing price appreciation has probably already vanished from the home equity line in housing balance sheets, and turned into consumption by retirees who have moved away from America's priciest areas. The Survey of Consumer Finances data show that home equity has risen much more slowly than aggregate housing wealth, because rising mortgage levels have offset rising home values. Younger Americans, in particular, are more likely to have paid for their homes using large mortgages than to have experienced large wealth increases.

Overall, these shifts in housing wealth seem to show that older groups in certain geographic areas are receiving most of the gains, but we have not established causality. More research is needed to identify causality, especially because nonhousing wealth is skewing in somewhat similar ways among the groups noted above.

Boom–bust housing cycles can be important redistributors of wealth, too. Pfeffer, Danziger, and Schoeni (2013) document that the median household in the Panel Study of Income Dynamics lost more than 50 percent of its wealth between 2007 and 2011, and that 83 percent of that loss came from real estate. Wolff (2014) found that

in 2010, 16.2 percent of homeowners under the age of 35 had negative home equity, but only 5.3 percent of homeowners between 55 and 64 had negative home equity.

Housing supply shapes these wealth transfers because it partially determines the extent of a housing convulsion. Glaeser, Gyourko, and Saiz (2008) show that the 1980s housing boom and subsequent bust largely bypassed places with elastic housing supply. In those years, buyers seem to have recognized that where it was easy to build, housing prices would not remain above construction costs for long. Consequently, the transfers of wealth that occurred during that boom were located primarily in places with restricted supply. The boom of the 2000s also disproportionately impacted places with limited supply, yet there were some areas such as Phoenix and Las Vegas that experienced booms despite enjoying relatively elastic housing supply. Because it takes time to build, overoptimistic buyers can still bid prices up in such markets for a few years. Eventually, the glut of new building in Las Vegas generated one of the largest of America's housing busts. Nonetheless, Mian, Rao, and Sufi (2013) show that wealth losses, and associated consumption declines, were higher in places where housing is less elastically supplied.

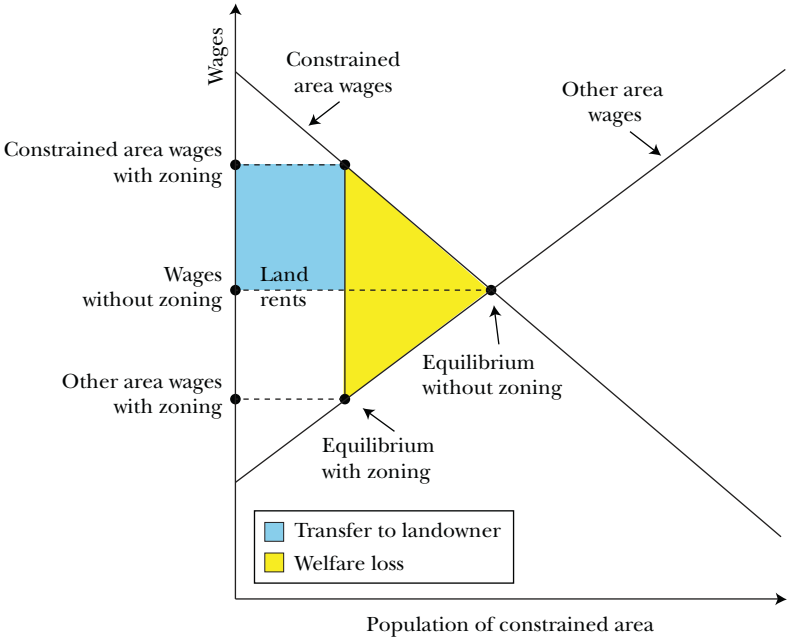
The Impact of Supply Restrictions: Urban Labor Markets and Productivity

Rising house prices represent a transfer from buyers to sellers, which is not itself obviously a welfare gain or loss. Yet constricted housing supply also generates a potentially profound distortion: people are unable to move into more desirable metropolitan areas. Hsieh and Moretti (2017) and Ganong and Shoag (2013) have raised the possibility that housing restrictions have led to a misallocation of labor that could have a serious adverse effect on US GDP. Given the large differences in productivity between Las Vegas and San Francisco, it seems virtually certain that America's GDP would rise if, for example, the San Francisco Bay Area built more housing, allowing more population to shift there from Las Vegas.

To better understand the possible GDP gains from eliminating land use controls, it is useful to make simplifying assumptions, some of which can bias the calculation in ways that are discussed at the end of this section. One such assumption is that there are no differences in negative externalities across locations. While there is little evidence to suggest that the negative effect of an extra home in a constrained area is worse than the negative effect of an extra home in an unconstrained area, if the externalities of construction were far worse in some places than others, then our estimates will overstate the benefits of deregulating housing markets.[10] Another is that construction costs are the same everywhere, which as discussed above is a

[10] Glaeser and Ward (2009) show that if one assumes constant construction costs (a rough but reasonable assumption, as discussed earlier), then land values are maximized when the gap between the mark-up over construction costs relative to price is equal to the absolute value of the elasticity of price with respect to density. Glaeser and Ward find that this gap is roughly ten times larger than the elasticity.

Figure 4
Welfare Consequences of Restricting Development in a Productive Market



roughly plausible assumption. We will also ignore amenity differences, so an absence of regulation will tend to equalize housing costs and wages across space.

In this setting, the potential output benefits from reallocating a fixed amount of labor from low-wage areas to high-wage areas can be seen in Figure 4, which depicts demand curves for two areas. The horizontal axis shows population in the constrained area, and a higher population in that area causes wages to decline. Population in the unconstrained area is the remainder, and so more population in the constrained area means less in the unconstrained area, leading to the upward-sloping demand curve for labor shown here. In the absence of land use controls, prices equalize across the two areas, which is shown in the point in the middle of the figure where the two curves meet. When housing supply is restricted, the wage in the restricted area is higher than in the unrestricted area.

If we assume that the demand for housing comes only from local labor markets, then we can treat each of these lines as a transformation of the labor demand curve, which in turn reflects the marginal product of labor. The lost output from misallocation is then equal to the area under the higher line from the restricted population level to the level that causes the lines to meet. This difference represents a classic deadweight loss triangle. In addition, there is a rectangle that represents the transfer to the owners of land in the more expensive area.

Hsieh and Moretti (2017) offer a set of illustrative calculations that have received considerable attention. They use a Cobb–Douglas production function in which the share of labor is 0.65 and the share of fungible capital—which will move in response to shifts in labor between cities—is 0.25. In this framework, the elasticity of labor demand is −7.5. In their analysis, changing the housing supply regulation in just three highly constrained markets—New York, San Francisco, and San Jose—to the median for the country results in a nearly 9 percent rise in aggregate GDP. This is achieved via massive shifts in employment location. Jobs in the New York market increase by 1,010 percent, with those in San Jose rising by 689 percent. Naturally, output is much higher in these markets, too. Wages in these areas do fall, but only by 25 percent in their model.

The Cobb–Douglas production function with fungible capital is an important driver of this result in which cities can grow enormously with relatively modest decreases in wages. Assumptions about the shape of the labor demand function also have a strong effect in shaping the conclusions about the welfare losses from distortions in labor supply. Cobb–Douglas production functions tend to deliver particularly elastic labor demand curves, especially when capital is also mobile. Consequently, they lead to the conclusion that even relatively small wage gaps will result in large population misallocations and welfare losses.

Empirical estimates of the link between wages and labor demand at the local level are often much lower than predictions from a Cobb–Douglas function. Beaudry, Green, and Sand (2014) present city-level labor demand elasticities that seem matched to our needs. They find a city-level labor elasticity of −0.3, which suggests that the overall impact is 0.7 percent of GDP. Their city-industry-level estimates are larger (−1.0), and those would imply a misallocation cost equal to about 2 percent GDP. Past demand elasticities have typically ranged from −0.25 to −1.0 (Hammermesh 1991). In addition, we have experimented with back-of-the-envelope estimates of these gains using linear demand functions for labor, rather than the curved demand functions implied by the Cobb–Douglas function. While the precise outcome depends on the parameters used, such calculations suggest that 2 percent of GDP may be an upper bound on the gains from the reallocation of labor.[11]

We view any gain that involves adding several percent to GDP as quite sizeable and worth pursuing. But clearly, considerable work remains to be done in pinning down the likely size of the potential gains. This follow-up work might also keep in mind the likely biases from our simplifying assumptions.

In empirically estimating the costs of labor misallocation, we also should be cognizant of the problem of omitted human capital. The average worker in Tulsa will not necessarily earn the average wages in Silicon Valley by moving to San Jose. Any misallocation calculation will typically increase with the variance in perceived productivities, and the noise created by unobserved human capital heterogeneity will generally cause an overestimate of misallocation costs.

[11] For examples of these calculations, see the online Appendix available with this paper at http://e-jep.org.

Another issue is that if places with higher human-capital-adjusted wages typically have more amenities (because cities are more likely to form in areas that are either productive or nice or both), then differences in the cost of housing will lead to an overestimate of the true differences in productivity. Conversely, there are some examples of large urban areas, like Orlando and Miami, that have lower-than-average wages and housing prices but also have the amenity of Florida sunshine. Again, not taking that amenity into account will bias attempts to infer productivity from wage levels.

On the other side, our calculations reflect only an estimate based on static factors. One might speculate that Silicon Valley and other high-productivity urban areas are about creativity, as well as high wages. If so, then, more Silicon Valley residents could also mean more technological innovation and faster productivity growth. If agglomeration economies are important, and tend to increase with population size, then this will attenuate the downward impact of added population on earnings. We are ignoring the impact that higher output has on product demand, which is captured in Hsieh and Moretti (2017), which also pushes earnings and the benefits from better labor allocation upward.

Next, the reallocation of population implied in this analysis would mean that the overwhelming majority of cities would lose population, while a few, such as New York and the San Francisco Bay Area, would gain substantial numbers of workers. In some of our back-of-the-envelope calculations, the entire population of certain cities would depart! As discussed earlier in the paper, declines in local demand for housing, given the durability of housing, can easily cause housing prices in those cities to fall—which further complicates calculations about what reallocation of population and welfare gains might be possible as a result of less-stringent limits on housing construction. Finally, we stress again that we have assumed away any benefits that regulation might create by reducing the negative externalities from construction, so these estimates should be taken as suggestive, not definitive.

Conclusion

When housing supply is highly regulated in a certain metropolitan area, housing prices are higher and population growth is smaller relative to the level of demand. While most of America has experienced little growth in housing wealth over the past 30 years, the older, richer buyers in America's most regulated areas have experienced significant increases in housing equity. The regulation of America's most productive places seems to have led labor to locate in places where wages and prices are lower, reducing America's overall economic output in the process.

Advocates of land use restrictions emphasize the negative externalities of building. Certainly, new construction can lead to more crowded schools and roads, and it is costly to create new infrastructure to lower congestion. Hence, the optimal tax on new building is positive, not zero. However, there is as yet no consensus about

the overall welfare implications of heightened land use controls. Any model-based assessment inevitably relies on various assumptions about the different aspects of regulation and how they are valued in agents' utility functions.

Empirical investigations of the local costs and benefits of restricting building generally conclude that the negative externalities are not nearly large enough to justify the costs of regulation. Adding the costs from substitute building in other markets generally strengthens this conclusion, as Glaeser and Kahn (2010) show that in America building restrictions are higher in places that have lower carbon emissions per household. If California's restrictions induce more building in Texas and Arizona, then their net environmental effect could be negative in aggregate. If restrictions on building limit an efficient geographical reallocation of labor, then estimates based on local externalities would miss this effect, too.

If the welfare and output gains from reducing regulation of housing construction are large, then why don't we see more policy interventions to permit more building in markets such as San Francisco? The great challenge facing attempts to loosen local housing restrictions is that existing homeowners do not want more affordable homes: they want the value of their asset to cost more, not less. They also may not like the idea that new housing will bring in more people, including those from different socioeconomic groups.

There have been some attempts at the state level to soften severe local land use restrictions, but they have not been successful. Massachusetts is particularly instructive because it has used both top-down regulatory reform and incentives to encourage local building. Massachusetts Chapter 40B provides builders with a tool to bypass local rules. If developers are building enough formally defined affordable units in unaffordable areas, they can bypass local zoning rules. Yet localities still are able to find tools to limit local construction, and the cost of providing price-controlled affordable units lowers the incentive for developers to build. It is difficult to assess the overall impact of 40B, especially since both builder and community often face incentives to avoid building "affordable" units. Standard game theoretic arguments suggest that 40B should never itself be used, but rather work primarily by changing the fallback option of the developer. Massachusetts has also tried to create stronger incentives for local building with Chapters 40R and 40S. These parts of their law allow for transfers to the localities themselves, so builders are not capturing all the benefits. Even so, the Boston market and other high-cost areas in the state have not seen meaningful surges in new housing development.

This suggests that more fiscal resources will be needed to convince local residents to bear the costs arising from new development. On pure efficiency grounds, one could argue that the federal government should provide such resources, but from a political economy perspective, the median taxpayer in the nation effectively transferring resources to much wealthier residents of metropolitan areas like San Francisco seems challenging to say the least. However daunting the task, the potential benefits look to be large enough that economists and policymakers should keep trying to devise a workable policy intervention.

■ *Edward Glaeser thanks the Taubman Center for State and Local Government at Harvard University for financial support. Joseph Gyourko thanks the Research Sponsor Program of the Zell/Lurie Real Estate Center at the Wharton School for financial support. The excellent research assistance of Yue Cao, Matt Davis, and Xinyu Ma is much appreciated, but the authors remain responsible for any errors.*

References

Barbosa, Filipe, Jonathan Woetzel, Jan Mischke, Maria Joao Ribeirinho, Mukund Sridhar, Matthew Parsons, Nick Bertram, and Stephanie Brown. 2017. *Reinventing Construction through a Productivity Revolution*. Report, McKinsey Global Institute, February, http://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/Reinventing-construction-through-a-productivity-revolution.

Beaudry, Paul, David A. Green, and Benjamin M. Sand. 2014. "In Search of Labor Demand." NBER Working Paper 20568, October.

Cheshire, Paul, and Steven Sheppard. 2002. "The Welfare Economics of Land Use Planning." *Journal of Urban Economics* 52(2): 242–69.

Davis, Morris A., and Jonathon Heathcote. 2007. "The Price and Quantity of Residential Land in the United States." *Journal of Monetary Economics* 54(8): 2595–2620.

Davis, Morris A., and Michael G. Palumbo. 2008. "The Price of Residential Land in Large US Cities." *Journal of Urban Economics* 63(1): 352–84.

Diamond, Rebecca. 2016. "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000." *American Economic Review* 106(3): 479–524.

Fishel, William A. 2001. *The Homevoter Hypothesis: How Home Values Influence Local Goverment Taxation, School Finance, and Land-Use Policies*. Cambridge, MA: Harvard University Press.

Frieden, Bernard J. 1979. "The New Regulation Comes to Suburbia." *Public Interest*, Spring 1979: 15–27.

Ganong Peter, and Daniel Shoag. 2013. "Why Has Regional Income Convergence in the U.S. Declined?" Harvard Kennedy School Working Paper No. RWP12-028, March 28.

Glaeser, Edward L., and Joseph Gyourko. 2003. "The Impact of Building Restrictions on Housing Affordability." *FRBNY Economic Policy Review* 9(2): 21–39.

Glaeser, Edward L., and Joseph Gyourko. 2005. "Urban Decline and Durable Housing." *Journal of Political Economy* 113(2): 345–75.

Glaeser, Edward L., and Joseph Gyourko. 2008. *Rethinking Federal Housing Policy: How to Make Housing Plentiful and Affordable*. Washington, DC: AEI Press.

Glaeser, Edward L., and Joseph Gyourko. 2017. "The Economic Implications of Housing Supply." Working Paper 802, Zell/Lurie Real Estate Center, January 4.

Glaeser, Edward L., Joseph Gyourko, and Albert Saiz. 2008. "Housing Supply and Housing Bubbles." *Journal of Urban Economics* 64(2): 198–217.

Glaeser, Edward L., Joseph Gyourko, and Raven Saks. 2005. "Why is Manhattan So Expensive? Regulation and the Rise in Housing Prices." *Journal of Law and Economics* 48(2): 331–69.

Glaeser, Edward L., and Matthew E. Kahn. 2010. "The Greenness of Cities: Carbon Dioxide Emissions and Urban Development." *Journal of Urban Economics* 67(3): 404–418.

Glaeser, Edward L., and Bryce A. Ward. 2009. "The Causes and Consequences of Land Use Regulation: Evidence from Greater Boston." *Journal of Urban Economics* 65(3): 265–78.

Gyourko, Joseph, Christopher Mayer, and Todd Sinai. 2013. "Superstar Cities." *American Economic Journal: Economic Policy* 5(4): 167–99.

Gyourko, Joseph, and Raven Molloy. 2015. "Regulation and Housing Supply." Chap. 19 in *Handbook of Regional and Urban Economics* 5B, edited by Gilles Duranton, J. Vernon Henderson, and William Strange. Amsterdam: Elsevier.

Gyourko, Joseph, and Albert Saiz. 2006. "Construction Costs and the Supply of Housing

Structure." *Journal of Regional Science* 46(4): 661–80.

Gyourko, Joseph, Albert Saiz, and Anita Summers. 2008. "A New Measure of the Local Regulatory Environment for Housing Markets: The Wharton Residential Land Use Regulatory Index." *Urban Studies* 45(3): 693–721.

Hammermesh, Daniel S. 1991. "Labor Demand: What Do We Know? What Don't We Know?" NBER Working Paper 3890, November.

Hayashi, Fumio. 1982. "Tobin's Marginal q and Average q: A Neoclassical Interpretation." *Econometrica* 50(1): 213–24.

Hseih, Chang-Tai, and Enrico Moretti. 2017. "Housing Constraints and Spatial Misallocation." NBER Working Paper 21154, May 18.

La Cava, Gianni. 2016. "Housing Prices, Mortgage Interest Rates and the Rising Share of Capital Income in the United States." BIS Working Paper 572, July.

Mian, Atif R., Kamalesh Rao, and Amir Sufi. 2013. "Household Balance Sheets, Consumption and the Economic Slump." Chicago Booth Research Paper 13-42.

Moretti, Enrico. 2013. "Real Wage Inequality." *American Economic Journal: Applied Economics* 5(1): 65–103.

Pfeffer, Fabian T., Sheldon Danziger, and Robert F. Schoeni. 2013. "Wealth Disparities Before and After the Great Recession." *Annals of the American Academy of Political and Social Science* 650(1): 98–123.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press.

Rognlie, Matthew. 2015. "Deciphering the Fall and Rise in Net Capital Share: Accumulation or Scarcity?" *Brookings Papers on Economic Activity*, Spring.

Rosenthal, Stuart, and William Strange. 2008. "The Attenuation of Human Capital Spillovers." *Journal of Urban Economics* 64(2): 373–89.

RSMeans. 2015. *RSMeans Building Construction Cost Data*, 73rd Annual Edition. Kingston, MA: RSMeans.

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

Sinai, Todd, and Nicholas S. Souleles. 2005. "Owner-Occupied Housing as a Hedge Against Rent Risk." *Quarterly Journal of Economics* 210(2): 763–89.

Turner, Matthew A., Andrew Haughwout, Wilbert van der Klaauw. 2014. "Land Use Regulation and Welfare." *Econometrica* 82(4): 1341–1403.

Wolff, Edward N. 2014. "Household Wealth Trends in the United States, 1962–2013: What Happened over the Great Recession?" NBER Paper 20733.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 31–58

Homeownership and the American Dream

Laurie S. Goodman and Christopher Mayer

For decades, it was taken as a given that an increased homeownership rate was a desirable goal. In May 1995, President Bill Clinton released the National Homeownership Strategy (US Department of Housing and Urban Development 1995), an 87-page, 100-point plan with the goal that it would "boost homeownership in America to an all-time high by the end of the century." President George W. Bush framed homeownership as a way to reduce racial inequality, and in 2003 signed the American Dream Downpayment Initiative to assist first-time homebuyers with obtaining a down payment (Bush 2003). But after the financial crises and Great Recession, in which roughly eight million homes were foreclosed on and about $7 trillion in home equity was erased, economists and policymakers are re-evaluating the role of homeownership in the American Dream. Many question whether the American Dream should really include homeownership or instead focus more on other aspects of upward mobility, and most acknowledge that homeownership is not for everyone.

In this article, we take a detailed look at US homeownership from three different perspectives. We first take an international perspective comparing US homeownership rates with those of other nations. The data show that the US homeownership rate is at the middle to lower end of the range relative to other developed countries.

■ *Laurie S. Goodman is Co-director, Housing Finance Policy Center, Urban Institute, Washington, DC. Christopher Mayer is the Paul Milstein Professor of Real Estate, Graduate School of Business, Columbia University, New York, New York, and a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are lgoodman@urban.org and cm310@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.32.1.31

doi=10.1257/jep.32.1.31

Moreover, the US rate is about the same as it was in 1990, while the homeownership rate has increased substantially in most other developed countries.

We then take a demographic perspective and examine the correlation between changes in the US homeownership rate between 1985 and 2015 and factors like age, race/ethnicity, education, family status, and income. The homeownership rate increased more in 1995 and 2005 and fell more in 2015 than can be explained by demographics. Part of the run-up in homeownership is likely due to relaxed credit standards and new mortgage products that expanded the borrower base and lowered default rates. Subsequently, in the aftermath of the Great Recession, homeownership fell with tight credit conditions, problematic student loan debt, stagnant real incomes, and perhaps a subtle change in attitudes toward homeownership. Racial and ethnic disparities in home ownership remain pronounced. Homeownership rates for black households have fallen every decade for the last 30 years, both unconditionally and after controlling for income and demographics. Even in 2015, black households with a college education are less likely to own a home than white households whose head did not graduate from high school.

Finally, we turn to the financial benefits of homeownership. Using national data since 2002, the internal rate of return to homeownership is quite favorable compared to alternative investments, even during a period where home prices suffered the worst shock since the Great Depression. While this result does not depend only on favorable tax treatment, tax subsidies certainly help increase the financial benefits of homeownership. Of course, these results vary with the timing of the purchase, the holding period, and location. Returns to homeownership have been less favorable in locations such as Cleveland and Chicago relative to metropolitan areas like Los Angeles, Dallas, and New York. We then consider other risks and benefits to homeownership not taken into account in our basic model. Homeownership does not seem to impair mobility across metropolitan areas during recessions. As well, homeownership appears to help borrowers accumulate housing and nonhousing wealth in a variety of ways, with tax advantages, greater financial flexibility due to secured borrowing, built-in "default" savings with mortgage amortization and nominally fixed payments, and the potential to lower home maintenance costs through sweat equity. However, the ability to build wealth through homeownership is dependent on holding on to the home during downturns; lower-income and minority borrowers are less likely to maintain homeownership through the cycle, and thus benefit less from homeownership.

Our overall conclusion: homeownership is a valuable institution. On average, it allows families to build wealth and serves as a measure of financial security. Homeownership rates in a variety of countries peak for households in their 60s, suggesting that owning a home helps reduce financial risk in retirement. Moreover, the mortgage interest deduction is not the main source of these gains; even if it were removed, homeowners would continue to benefit from a lack of taxation of imputed rent and capital gains, which are tax benefits available in most countries around the world. There are very substantial variations in the homeownership experience, depending on factors like purchase timing, holding period, and location. But while two decades of policies in the 1990s and early 2000s may have put too

Table 1
Global Homeownership Rates by Country and Year, 1990–2015

	Homeownership rate (percent)					*Change in homeownership rate*		
	1990	*2000*	*2005*	*2010*	*2015*	*1990–2005*	*2005–2015*	*1990–2015*
Bulgaria	89.8	96.5	85.4	86.9	82.3	–4.4	–3.1	–7.5
Canada	62.6	65.8	67.1	69.0	67.0	4.5	–0.1	4.4
Czech Republic	38.4	47.0	73.5	78.7	78.0	35.1	4.5	39.6
Denmark	54.5	51.0	66.6	66.6	62.7	12.1	–3.9	8.2
Finland	67.0	61.0	71.8	74.3	72.7	4.8	0.9	5.7
France	54.4	54.8	61.8	62.0	64.1	7.4	2.3	9.7
Germany	37.3	41.3	53.3	53.2	51.9	16.0	–1.4	14.6
Ireland	80.0	78.9	78.2	73.3	70.0	–1.8	–8.2	–10.0
Italy	64.2	69.0	72.8	72.6	72.9	8.6	0.1	8.7
Japan	63.2	64.9	63.1	62.4	64.9	–0.1	1.8	1.6
Mexico	78.4	72.7	71.3	69.8	71.7	–7.1	0.4	–6.7
Singapore	87.5	92.0	91.1	87.2	90.8	3.6	–0.3	3.3
Slovenia	68.0	82.3	83.2	78.1	76.2	15.2	–7.0	8.2
Spain	77.8	82.0	86.3	79.8	78.2	8.5	–8.1	0.4
Sweden	41.0	67.0	68.1	70.8	70.6	27.1	2.5	29.6
Switzerland	31.3	34.6	38.4	44.4	51.3	7.1	12.9	20.0
United Kingdom	65.8	69.1	69.2	65.7	63.5	3.4	–5.7	–2.3
United States	**63.9**	**66.8**	**68.9**	**66.9**	**63.7**	**4.9**	**–5.2**	**–0.3**
Average	**62.5**	**66.5**	**70.6**	**70.1**	**69.6**	**8.1**	**–1.0**	**7.1**

Notes: Due to differing census and survey years, many figures in the table are from a year or two before or after the listed year, or the average between two nearby values. Sources for individual countries are listed in the Data Appendix.

much faith in the benefits of homeownership, the pendulum seems to have swung too far the other way, and many now may have too little faith in homeownership as part of the American Dream.

Homeownership around the World

The United States does not rank particularly high among other high-income countries when it comes to homeownership. Table 1 compares the homeownership rate from 1990 to 2015 across 18 countries where we have been able to obtain somewhat comparable data over the entire time period. The United States was ranked tenth in 1990, at the middle of the pack and close to the mean rate. By 2015, the United States was the fifth-lowest, its homeownership rate of 63.7 percent falling well below the 18-country average of 69.6 percent. Over the 1990–2015 period, 13 of the 18 countries increased their homeownership rates. The five countries with declines in homeownership were Bulgaria, Ireland, Mexico, the United Kingdom—and the United States.

In a broader sample of countries, many of which have missing data for some of the years in question, the United States homeownership rate in 1990 was slightly below the median and mean of the 26 countries reporting data. By 2015, the US

ranked 35 of 44 countries with reliable data, and was almost 10 percentage points below the mean homeownership rate of 73.9 percent. In the online appendix Table A1-1 (available with this paper at http://e-jep.org), we report results that include an additional 30 countries. We also give a couple of data sources.

By contrast, the age-pattern of homeownership in the United States is similar to that of other European countries. In most countries, homeownership rates peak at or near retirement, between ages 65 to 74. Other than Germany, Austria, and the Netherlands, the homeownership rate at this age peaks between 75 and 90 percent (it is 80 percent in the United States), well above the rate for younger households. Home equity for seniors in large European countries exceeds 8 trillion euros in 2013 (compared to over 5 trillion euros in the United States). This pattern suggests that home equity often plays an important role in retirement savings, although homeowners often don't access the equity directly except through the rent-free use of the property.[1]

Looking at the reasons behind differences in homeownership across countries can be difficult. Each country has its own culture, demographics, policies, housing finance systems, and, in some cases, a past history of political instability that favors homeownership (Butrica and Mudrazija 2017). Badarinza, Cambell, and Ramadorai (2016) offer evidence on differences in household balance sheets for 13 countries and a discussion of various institutions such as the mortgage markets across these countries. The authors point to a linkage between mortgage finance, pensions, equity participation, and homeownership. While not definitive, countries like France, Germany, and the Netherlands have both lower-than-average homeownership rates and robust public pensions and private defined-contribution systems.

As well, government tax policy and regulations appear to play an important role in countries with below-average homeownership rates. For example, consider the evolution of homeownership in (the former) West Germany and the United Kingdom (Phillips 2014). Both countries pursued a similar policy of subsidizing postwar rental construction to rebuild their countries. However, in intervening years, German policies allowed landlords to raise rents to some extent and thus finance property maintenance while also providing "protections" for renters. In the United Kingdom, regulation strongly discouraged private rentals, whereas the quality of public (rental) housing declined with undermaintenance and obtained a negative stigma. As well, German banks remained quite conservative in mortgage lending. The result was that between 1950 and 1990, West German homeownership rates barely increased from 39 to 42 percent, whereas United Kingdom homeownership rates rose from 30 to 66 percent. Interestingly, anecdotes suggest that many German households rent their primary residence, but purchase a nearby home to rent for income (which requires a large down payment but receives generous depreciation benefits). This allows residents to hedge themselves against the potential of rent increases in a system that provides few tax subsidies to owning a home.[2]

[1] For further detail, see Figure A1-2 in the online Appendix as well as Haurin and Moulton (2017).

[2] We thank Michael Lea, Deborah Lucas, and Mark Zandi for their helpful comments on the details of the German housing finance system.

Switzerland also has a low homeownership rate, and once again, tax regulations favor renting over owning. Bourassa and Hoesli (2010) conclude that income tax policy, especially the tax on imputed rents, as well as the high price of owning relative to renting are key determinants of why many more Swiss households are renters than in other countries. On the other side of the equation, the Netherlands, Switzerland, and the United States all have relatively generous mortgage interest deductions.

Patterns in US Homeownership Rates

The overall US homeownership rate rose from 63.5 percent in 1985 to 65.0 percent in 1995 and peaked at 68.8 percent in 2005. It then dropped to 62.7 percent by 2015, according to data from the American Housing Survey. We argue that neither the rise nor the fall of the homeownership rate can be explained by demographic changes alone, like the population becoming older or better educated. Rather, we argue, the vast expansion in credit contributed to the rise in the homeownership rate from 1985 to 2005, and the effects of the Great Recession, in combination with student loan debt, tight credit, and a subtle change in attitudes toward homeownership contributed to the fall in homeownership from 2005 to 2015. Homeownership rates for blacks have declined relative to whites and Asians, a fact that cannot be easily explained by household income or demographics.[3]

Demographic Factors Contributing to Homeownership

Table 2 shows the homeownership rate by race/ethnicity, age, education, and household composition. With a few exceptions, which we discuss below, the homeownership pattern across groups is the same: it increases from 1985 to 2005, then falls dramatically between 2005 and 2015.

Several demographic patterns in the table have implications for patterns of ownership over time. For example, the homeownership rate increases with age, peaking during retirement age after 65. After 1985, the homeownership rate for the 85+ group is consistently higher than for those who are 35 to 44. Over time, the US population has become older. For example, the share of households in which the head was 44 or younger fell from 49.2 percent in 1985 to 35.7 percent in 2015; conversely, the share of households in which the head was 65 or over rose from 21.5 percent in 1985 to 23.9 percent by 2015 (for details, see Table A-2.1 of the online Appendix). An aging population should contribute to a rising homeownership rate.

[3]The most commonly cited measure of homeownership comes from the Current Population Survey as reported by the US Census Bureau. However, for this current paper, we have chosen to use data from the American Housing Survey, which is a nationally representative longitudinal survey conducted every two years. The AHS data closely mirror the CPS data in overlapping years, but the AHS provides additional detail on households and housing units. The AHS has been conducted in a similar format since 1985, although in 2015 a new sample was selected and some reported variables changed. We were reluctant to estimate using the decennial census for the back data or gather more recent data since 2010 from another dataset like the American Community Survey, as the two series are not totally consistent.

Table 2
Homeownership Rates

	1985	*1995*	*2005*	*2015*
Overall	63.5%	65.0%	68.8%	62.7%
Race				
White	68.3%	71.4%	75.8%	70.8%
Black	43.9%	43.6%	48.5%	42.2%
Asian, Pacific Islander	45.0%	53.2%	61.1%	56.6%
Hispanic	39.6%	41.8%	49.4%	45.4%
Other	44.1%	43.1%	53.8%	49.0%
Age				
15–24	16.5%	14.2%	23.9%	10.8%
25–34	45.5%	45.4%	49.2%	34.5%
35–44	68.0%	65.5%	68.7%	56.4%
45–54	75.2%	75.5%	76.7%	67.3%
55–64	79.3%	79.4%	81.1%	74.8%
65–74	77.6%	81.3%	82.8%	78.9%
75–84	68.1%	76.9%	80.9%	79.0%
85 +	60.8%	66.1%	68.9%	70.7%
Education level				
Less than high school	61.0%	58.2%	57.1%	48.6%
High school	63.8%	65.4%	68.2%	60.4%
Some post-secondary	60.9%	67.5%	72.3%	63.9%
College degree or higher	68.1%	71.8%	76.7%	71.4%
Household Composition				
Living alone, male	37.5%	42.1%	50.6%	48.8%
Living alone, female	51.5%	54.5%	59.4%	54.1%
Married couple with kids	73.7%	76.0%	79.1%	70.8%
Married couple without kids	81.5%	84.0%	87.2%	82.5%
Male single householder, with kids	48.4%	53.0%	52.6%	45.6%
Male single householder, no kids	41.8%	45.1%	49.5%	46.2%
Female single householder, with kids	34.3%	38.5%	42.5%	32.8%
Female single householder, no kids	53.6%	57.7%	59.5%	52.6%

Source: American Housing Survey, 1985, 1995, 2005, and 2015.

Broadly speaking, all age groups saw their homeownership rate peak in 2005, but households in the prime home-buying ages of 35–54 saw less than a 1.5 percentage point increase in homeownership over the 20 years prior to 2005. Instead, the largest increases in homeownership were for households whose heads were 65–84, which was predominantly driven by cohorts whose income and wealth substantially increased in their working years (Mayer 2017). Thus, much of the increase in homeownership between 1985 and 2005 was driven by a large cohort of retirees whose homeownership rate was much higher than the previous cohort of retirees, while homeownership rates of households in prime home-buying years were relatively flat until the last decade, when they fell sharply after the Great Recession. The younger the age group, the sharper the decline in homeownership by 2015.

Those with more education are more likely to be homeowners, as shown in Table 2. Educational levels have also risen over time: from 1985 to 2015, the share of household heads with a high school or less education fell from 61.3 to 44.6 percent,

while the share of household heads who are college graduates rose from 21.5 to 39.8 percent. This pattern should also increase the homeownership rate.

In 1985, homeownership rates were broadly similar for all education groups, with only 7.1 percentage points separating households whose head does not have a high school degree (61.0 percent) from those with a college degree (68.1 percent). This relatively egalitarian pattern has sharply changed. By 2015, there was about a 23-percentage point difference in the home ownership rates of the most (71.4 percent) and least (48.6 percent) educated households. The decline in homeownership for those with a high school education or less is an especially striking pattern. As has been repeatedly pointed out in academic research, the least-educated workers have faced flat or falling real incomes and lower labor force participation in recent decades (Cynamon and Fazzari 2014; Gordon 2012; Aaronson and Mazumder 2005).

Hispanics and non-whites have considerably lower homeownership rates than their non-Hispanic white counterparts (hereafter referred to as "white"), as shown in Table 2. Moreover, the changes over the 1985–2015 period have been uneven, with white homeownership increasing by 2.5 percent, Hispanic homeownership increasing by 5.8 percent, Asian homeownership increasing by 11.6 percent, and black homeownership declining by 1.7 percent. While some portion of the racial and ethnic differences in homeownership is driven by socioeconomic variables, regression analysis shows that a substantial gap remains. In fact, the homeownership rate in 2015 was higher for whites with less than a high school education (62.9 percent) than for blacks with a college education (57.4 percent). The United States is becoming more racially/ethnically diverse: in 1985, 81 percent of the population was white, this declined to 67.1 percent by 2015 (for details, see Table A-2.1 in the online appendix). All things being equal, the increase in household diversity should have put a drag on the homeownership rate over the 1985–2015 period. But other factors have not remained constant: for example, the differences in wealth by educational attainment have increased considerably (McKernan, Ratcliffe, Steuerle, and Zhang 2013; Urban Institute 2015).

Married couples are much more apt to be homeowners than either those living alone or single householders living with other relatives; the percentage of households consisting of married couples declined from 57.3 percent in 1985 to 49 percent in 2015. Married couples with at least one child under age 18 were the single largest household category in 1985, describing 28.8 percent of households. By 2015, however, only 19.7 percent of the households fit into this category. There are now considerably more married households without children than with children. Homeownership declined for all types of households with children between 1985 and 2015, whether or not headed by a married couple.[4]

Clearly, demographics have exerted various pushes and pulls over homeownership in recent decades. In the next section, we offer a descriptive regression of

[4]While an earlier literature suggested that homeownership benefitted the children of homeowners (Dietz and Haurin 2003), more recent papers have suggested that this effect was largely due to selection and finds few differences in outcomes for children regardless of the tenure choice of their parents (Barker and Miller 2009; Holupka and Newman 2012).

these factors. Or course, the goal of this analysis is not to determine causality, but rather to summarize patterns that can be compared to previous research and may be further explored in future analysis. Along with the demographic variables, we use year dummy variables, which allows us, in each survey year, to estimate the size of homeownership changes that cannot be explained by observed demographics.

A Regression Illustration

Our regression approach is similar to that of Schwartz, Bostic, Green, Reina, Davis, and Augustine (2016), who study patterns affecting rental housing using factors that have been established to be important in previous research (Herbert, Harin, Rosenthal, and Duda 2005; Haurin and Rosenthal 2007). We use American Housing Survey data from 1985, 1995, 2005, and 2015. Our approach is to use a series of dummy variables so that, in each broad category, the coefficient should be interpreted as relative to the left-out variable.

Table 3 shows the regression results. In general, the coefficients are as expected. The first group of dummy variables reflect race/ethnicity of head of household, and the coefficients should be interpreted as compared to the left-out category of "White." Even controlling for income, education, age, and household type, homeownership rates vary substantially by race and ethnicity. Blacks, Hispanics, and Asians all had lower homeownership rates than their white counterparts. We experimented with some other control variables (described below), which reduce but do not eliminate this difference, suggesting that other factors beyond this analysis drive racial/ethnic differences in homeownership.

Previous research has consistently found that regressions do not explain black/white differences in owning a home. For example, Charles and Hurst (2002) points to smaller down-payment assistance from relatives and a higher likelihood of mortgage rejection as additional factors that contribute to lower homeownership rates for blacks, but still find a significant gap in the willingness of blacks to apply for a mortgage relative to whites. Haurin, Herbert, and Rosenthal (2007) suggest other additional factors may also play a role in the homeownership gap, including higher income volatility for blacks, lower family wealth, and differences in the neighborhoods where blacks are more likely to live. Bond and Eriksen (2017) find that 65 percent of the homeownership gap between blacks and whites can be explained by adding parents' attributes like wealth and whether they were homeowners in addition to other typical demographic and income variables. Indeed, because household wealth is not accurately captured on a mortgage application, and family wealth is certainly not captured, these regression results will overstate racial differences.

Nonetheless, research does not yet fully explain why blacks have persistently lower homeownership rates, or why this gap (after adjusting for other factors) has increased. Racial discrimination in some form is a possible explanation for the persistent white/black gap in homeownership. However, given the large amount of resources that policymakers have placed into closing the gap in lending by race of borrower and neighborhood demographics, it seems unlikely that the larger white/black gap in homeownership is being driven by a rise in discrimination alone.

Table 3

Results of a Regression Investigating the Relationship between Various Demographic Factors and Homeownership

Intercept	0.66628***	(< 0.0001)
Non-Hispanic black	–0.15330***	(< 0.0001)
Hispanic	–0.18876***	(< 0.0001)
Asian/Pacific Islander	–0.15455***	(< 0.0001)
Other race	–0.14127***	(< 0.0001)
log of household income	0.02976***	(< 0.0001)
Aged 15–24	–0.56348***	(< 0.0001)
Aged 25–34	–0.38944***	(< 0.0001)
Aged 35–44	–0.22215***	(< 0.0001)
Aged 45–54	–0.12445***	(< 0.0001)
Aged 55–64	–0.04940***	(< 0.0001)
Aged 75–84	0.00685	(0.149)
Aged 85 or more	–0.03263***	(< 0.0001)
Less than high school	–0.10006***	(< 0.0001)
High school graduate	–0.04492***	(< 0.0001)
Some postsecondary	–0.01929***	(< 0.0001)
1995	0.02501***	(< 0.0001)
2005	0.05808***	(< 0.0001)
2015	–0.01427***	(< 0.0001)
Male living alone	–0.25886***	(< 0.0001)
Female living alone	–0.23834***	(< 0.0001)
Married, with kids	0.06418***	(< 0.0001)
Single male (kids/no kids)	–0.16952***	(< 0.0001)
Single female, with kids	–0.20112***	(< 0.0001)
Single female, no kids	–0.16962***	(< 0.0001)
R^2	0.260	

Source: Authors using American Housing Survey data from 1985, 1995, 2005, and 2015.

Note: The table shows the results of a regression investigating the relationship between various demographic factors and homeownership. For each each category of dummy variables, the coefficient should be interpreted as relative to the left-out variable. The first group reflects race/ethnicity of head of household, with the omitted category being "White." For household head age groups, the omitted group is "Aged 65–74." For education of the household head, the omitted variable is "College." The omitted variable for household type, is "Married, no children." Finally, the year 1985 is omitted for the year dummy variables. Standard errors are shown in parentheses. *,**, and *** indicate significance at the .1, .01, and .001 levels respectively.

Other control variables have the expected sign and significance. We include (log of) household income as a control variable, and it has a strong, positive correlation with homeownership. Age groups are also included, with the omitted group being "Aged 65–74." With these control variables, the group with highest home ownership is aged 75–84, and the homeownership rate of the 85+ group is above that of the 55–64 age group. For education, the omitted variable is "College education."

Adding control variables does not eliminate the impact of education on homeownership. Relative to those with a college education, households whose heads have a lower educational level are less likely to be homeowners.

The base household type, that is, the omitted variable for household type, is "Married, no children." Not surprisingly, married households with children have the highest homeownership rate. All other (unmarried) household types have lower homeownership rates.

Finally, the year 1985 is omitted for the year dummy variables, and so the other coefficients show that even after adjusting for the other factors included here, the homeownership rate was 2.5 percent higher in 1995 than in 1985, 5.8 percent higher in 2005 than in 1985, and 1.4 percent lower in 2015 than in 1985. Thus, homeownership rates adjusted for the other demographic factors given here fell by a striking 7.3 percentage points from 2005 to 2015. Despite the reasonably large number of controls in these regressions, the size of the change in the year dummies is quite similar to the aggregate changes in homeownership rates, which suggests that most of the changes in homeownership are not being driven by the changes in the demographic variables. These are largely offsetting, with the rising age and education having a positive effect, and the increasingly non-white population and fewer families with children having a negative effect. Rather, the change in the homeownership rate is being driven by changes in the external environment, a point we will return to below.

Alternative Specifications

We experimented with a range of other specifications of the basic regressions, and the results are available in the online Appendix. While overall the results are qualitatively quite similar, we want to call attention to a few points.

In one specification, we used a more flexible indicator for income: specifically, using both a term for income and for income-squared. The greater flexibility for the income variable substantially reduces the magnitude and statistical significance of the education variables, which (not surprisingly) is consistent with the belief that the predominant impact of education on homeownership is via earnings.

In another specification, we ran four regressions, one for each quartile of income. While a household that is married with children generally has a higher homeownership rate (versus married without children), that is not the case in the lowest quartile, where the homeownership rate is unrelated to the presence of children in the household. Whatever the aspiration to become a homeowner, it is surely harder to save for a down payment when a household with low income must also support children. The coefficient of the income variable is also very different across quartiles. In the bottom quartile, the coefficient on income is quite small and negative, possibly suggesting the impact of homeownership programs that are targeted to the lowest income households. Again, this result is not surprising. The coefficient on income is quite high for the middle two quartiles, where incremental earnings may make a big difference in saving for a home and supporting a mortgage

payment. Income has a much smaller impact on homeownership for households in the top quartile.

What Factors Caused the Changes over Time?

Demographic factors underpredict the homeownership rate in 1995 and 2005, according to the year dummy variables, but overpredict it in 2015. Why is this? A number of factors seem to be at work.

The run-up in the homeownership rate from 1995 to 2005 can be partially explained by the emergence of nontraditional products and relaxation of credit standards, expanding the number of borrowers who could qualify. Mian and Sufi (2009, 2014) argue that the increase in mortgage credit was unrelated to fundamentals like income growth or lender expectations of house price appreciation, and indeed was not related to demand-side fundamentals, but instead to the supply of credit through the increase in securitization. For example, Mayer, Pence, and Sherlund (2009) calculate that 6.8 million subprime and Alt-A loans were originated between 2003 and 2005, and of those, about 2.8 million were purchase loans (as opposed to refinancing of existing mortgages). If half of those new purchase loans were for buyers who would not have been able to purchase without obtaining a nontraditional mortgage product, the homeownership rate would have risen about 1.6 percentage points, all else equal, or almost one-half of the 3.3 percent increase in the homeownership rate between 1995 and 2005 (see the coefficients for those years in Table 3).

Others have argued that demand for homeownership grew as household expectations that home prices would appreciate increased demand for owner-occupied properties. (Foote, Gerardi, and Willen 2012). In addition, the relatively rapid rise in home prices in many areas during the 1985–2005 period contributed to a low realized default rate, ensuring that even households facing financial challenges were able to maintain homeownership and lenders were more comfortable expanding credit (Gerardi, Lehnert, Sherlund, and Willen 2008).

To explain the decline in the homeownership rate between 2005 and 2015, there are at least four factors largely unrelated to demographic changes: the effects of the Great Recession, student loan debt, tight credit, and a shift in attitudes toward homeownership. Goodman, Pendall, and Zhu (2015) discuss these elements in greater detail and point out that they are difficult to separate empirically.

We can try to calculate the direct effect of the Great Recession on the homeownership rate. Hope Now (2017) (an organization including government, housing advocates, mortgage industry members, and investors) estimates there were, cumulatively, nearly eight million liquidations from the third quarter of 2007 to the end of 2015. We don't know how many of these were owner-occupied, as many investment property borrowers claimed to be owner-occupied. Assuming that six million of these were owner-occupied, and that under normal circumstances, two million owner occupied borrowers might have suffered a foreclosure over a similar time period, the incremental four million liquidations contributed to a roughly 3.3 percent drop in the home ownership rate (that is, 4 million additional owner-occupied foreclosures divided by 120 million households).

The amount of student loan debt has increased dramatically and likely contributed to a decline in the homeownership rate, especially for those who accumulated student debt but then did not graduate with a BA degree. From 2005 to 2015, the number of borrowers with student loan debt increased from 24.0 million to 43.3 million and the student loan debt balances grew from $378 billion to $1.19 trillion, according to the Federal Reserve Bank of New York's Consumer Credit panel. However, 41 percent of those starting college fail to complete their degree within 6 years (as reported at https://nces.ed.gov/fastfacts/display.asp?id=40). Gicheva and Thompson (2015) and Allison (2015) show that student loan debt is primarily an issue for those who do not receive their degree. For those who graduate, higher income offsets the impact of the debt and there is no net effect on homeownership.

Tight credit in the aftermath of the financial crises has also taken its toll on the homeownership rate. Li and Goodman (2014 with updates) look at the expected probability of default taken by the market in each origination quarter and show that the market in 2015 was taking less than half the expected credit risk it took in 2001. When comparing 2015 to 2001, new and existing home sales were down 4 percent but mortgage applications were down 32 percent. In 2001, 30 percent of borrowers had credit scores less than 660; in 2015, only 10 percent (Goodman, Zhu, and Bai 2016).

Commentators have debated whether there has been a change in attitudes with respect to homeownership. Homeownership clearly remains an aspiration for the vast majority of households. A National Association of Realtors (National Association of Realtors 2017) survey asked non-homeowners if they wanted to become a homeowner in the future: 86 percent said "yes," a percentage that has been roughly constant through the years. A Fannie Mae survey (2014) asked younger renters if they plan to buy, and 90 percent said they will, eventually. However, in such survey data, the questions do not put a timeframe on the purchase or take into account the difference between aspiration and ability. A recent Freddie Mac (2017) survey found that even though renters are more optimistic about their financial situation, 59 percent said their next home would be a rental, up from 55 percent six months earlier. Moreover, of the 80 percent of renters that said they would like to own a home at some point, only 29 percent said they could afford to purchase now, 38 percent said they cannot afford to purchase now, and 14 percent said while they would like to own, they do not think they would ever be able to afford it.

Perhaps the best documentation of a change in willingness to become a homeowner comes from a Fannie Mae study in which Simmons (2014), used American Community Survey data in the aftermath of the financial crisis. After controlling for race/ethnicity, they found a much lower homeownership rate for 30–32 year olds who were married with at least one child in the home and at least $95,000 in income. That is, when looking at a sample of those who historically would have had a high desire along with the ability to purchase a home (and would not have been much affected by tight credit markets), there has been a marked decline in the percentage who actually purchased a home.

Notice that while we have looked only at national homeownership rates, there is a huge variation across the nation, with some states, particularly in the middle of the country, having much higher rates than others. There is also a difference

between metropolitan and nonmetropolitan areas, with non-metro areas generally having higher homeownership rates. Finally, certain expensive cities on the coasts have homeownership rates that are lower than both their state and other metro areas. Explaining this variation is a promising topic for future study.

Going forward, while some factors, like tight credit markets and borrowers who lost their homes in the aftermath of the Great Recession, may correct themselves, other challenges like higher student loan debts and labor market difficulties for low-income households are likely to persist. As a result, the relatively low homeownership rate in 2015 may stay low for an extended period of time.

Does Owning a Home Make Financial Sense?

A potential homeowner must consider a number of tradeoffs. We start by computing the financial returns, including tax benefits, associated with purchasing a home in 2003 relative to renting using estimates of sale prices and rents for the same homes. (Single-family homes for rent represent 13 percent of the housing stock, up from 9 percent a decade ago, as reported in Garrison 2015.) In the next section, we examine the nonfinancial costs and benefits.

Our results suggest that there remain very compelling reasons for most American households to aspire to become homeowners. Financially, the returns to purchasing a home in a "normal" market are strong, typically outperforming the stock market and an index of publicly traded apartment companies on an after-tax basis. Of course, many caveats are associated with this analysis, including variability in the timing and location of the home purchase, and other risks and tradeoffs associated with homeownership. There is little evidence of an alternative savings vehicle (other than a government-mandated program like Social Security) that would successfully encourage low-to-moderate income households to obtain substantial savings outside of owning a home. The fact that homeownership is prevalent in almost all countries, not just in the United States, and especially prevalent for people near retirement age, suggests that most households still view homeownership as a critical part of a life-cycle plan for savings and retirement.

Financial Returns to Buying a Home: The Framework

For a homeowner, a home is both a place to live and an investment. Under certain conditions, the net present value of the cash flows from owning a home, versus renting for a given holding period and investing the down payment, should be the same. These conditions include: no uncertainty about home prices and rents; a deterministic rate of inflation which affects both home prices and rents; no tax advantages to home ownership; known costs of home maintenance, property taxes, and insurance; no difference between home price appreciation, mortgage rates, and returns on other investments; a known holding period, and zero transactions costs to move between the purchase and rental decision. Of course, in the real world, with uncertainty, liquidity constraints, mobility costs, moral hazard, and many other factors, it is not surprising that people may prefer

ownership over rental, or vice-versa. Our approach will be to first compute returns to owning versus renting in a simple framework that ignores such factors affecting the household's decision to buy or rent a home. Then we discuss these factors in a following section.

Financial Returns to Buying a Home: Data

While the broad framework seems straightforward, comparing the financial returns of owning and renting requires quite a bit of data from different sources. One key challenge in this analysis is determining the market value of the *use* of the home for an owner-occupant, because no (readily available) data show rents and prices for the same properties. Given large quality differences in the typical rental apartment and owner-occupied home, just comparing apartment rents to single-family home prices may introduce appreciable errors. Instead, we rely on newly available Zillow data on median home prices and rents that are estimated for all the properties in its coverage universe. Zillow calculates an estimated home value ("Zestimate") and a separate estimated rent value at the property level using data on all rents and transactions in their database, and then takes the median. In theory, this approach should control for biases associated with differences between rental and owner-occupied homes and for variation in the types of properties selling over time. However, we do not have access to Zillow's proprietary model, and thus cannot examine the possibility of a changing value of attributes for rental versus owner-occupied properties or estimation errors that might be systematically biased.[5] Zillow provides data at the metropolitan area and for the nation as a whole.

Information on annual costs for homeowners are obtained from the American Housing Survey, which asks detailed questions about costs for homeowners (and renters). We use this data for costs of maintenance and capital improvements, and for property taxes at the national level. Because the American Housing Survey (AHS) is conducted every other year at the national level, we interpolate values for the middle years. After 2013, the AHS no longer reports detailed costs, and so we index the most recent values using the Consumer Price Index (CPI). At the metropolitan area level, we use state property tax estimates from the Lincoln Land Institute and Minnesota Center for Fiscal Excellence starting in 2006; for years 2003–2005, we used an annual property tax survey conducted by the District of Columbia.

[5] Zillow data are used in many academic projects due to their easy availability and perceived accuracy. Zillow reports a median error rate of 4.3 percent as of August 2017, meaning that half of all homes sell for a price within 4.3 percent of the current Zestimates. For more information, see https://www.zillow.com/zestimate/. For more detail on the Zillow methodology, see https://www.zillow.com/research/one-more-advance-in-creating-a-better-price-to-rent-ratio-2968/. The rental data are based on asking rents and may overstate rents at times of excess vacancies when landlords offer concessions. Given the strong demand for rental properties over this time period, asking rents are likely to be a good proxy for effective rents. Because Zillow only started publishing rents on its universe of properties in 2010, they provided us rental data that were indexed back to 2006 using data from the American Community Survey (ACS) at the metropolitan level. For prior years, Zillow used state-level median reported rent growth from the decennial censuses. The sample includes all properties in the Zillow database including single-family homes, condominiums, and cooperatives.

We also wish to compute the financial returns to purchasing a home relative to the returns from comparable indexes of alternative investments. Our analysis assumes a purchase at the end of 2002, a time when home prices were close to a long-run normalized level and prior to the large run-up in home prices from 2003–2006 and the subsequent decline from 2007–2012. We compare returns for each year of ownership with potential sales from 2004–2016 using a representative (median) home in the United States and then for a selected set of metropolitan areas.

Results for a Homebuyer in 2002

Table 4 reports results from our computations for the financial returns from owning a median home purchased at the end of 2002. As shown in the first column, the home is purchased in this example at a price of $134,200, with a down payment of 20 percent. The format is similar to the standard pro forma used in commercial real estate to assess the returns from an investment.

The analysis starts with the value of the use of the home as measured by the rent a homeowner would pay to live in a comparable property. This is similar to the concept of implicit rental income. Then we subtract the operating costs, including maintenance costs, property taxes, and homeowner's insurance, to obtain the equivalent of net operating income: the net financial benefit of living in a home before the impact of capital expenditures, taxes, and financing. This analysis ignores items like utility costs that would commonly be paid by residents whether they were owners or renters.

Next, we subtract annual capital expenditures and financing costs—in this case, yearly mortgage payments. This yields the imputed annual cash flow from living in the home. This annual imputed cash flow is negative for the first six years of ownership, which occurs in this example predominantly because the homeowner has chosen to use relatively high leverage of 80 percent, and the initial mortgage payment is 64 percent of the initial imputed rental cost. In this example, we assume the borrower refinances once, in 2012, reducing the mortgage interest rate by over 200 basis points, acknowledging that the refinance option is a contributor to the financial return on equity (Nothaft and Chang 2004). Of course, it is possible that many homeowners substituted "sweat equity" for cash capital expenditures during a time period when wage growth was low and thus our estimates of financial returns might not correctly incorporate the value of their labor (Bogdon 1996). Alternatively, homeowners might not have fully maintained their homes, leading to below-average appreciation rates over this time period for existing homes.

Finally, we estimate the value to an owner of taxes saved from deducting mortgage interest, property taxes, and some financing costs. Including tax savings, imputed cash flow is always positive. We report returns with and without the tax savings, because an estimated 40 percent of homeowners do not itemize deductions on their tax form (Lu and Toder 2016) and thus are not able to achieve this tax savings.

The next rows report financial cash flows for the purchase in December 2002, as well as the net sales proceeds for each year as if the owner sold the home between 2004 and 2016. This allows us to compute the internal rate of return (IRR) on a sale in any given year. The IRR is computed using the cash at purchase in 2002, the annual imputed cash flow for each year of ownership, and the cash at sale in

Table 4

Financial Returns from Owning a Home: National Data

Year	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*
Imputed rental "income" (Zillow)	**$11,611**	**$12,031**	**$12,452**	**$12,872**	**$13,293**	**$13,737**	**$14,294**	**$14,724**	**$14,986**	**$15,015**	**$15,021**	**$15,360**	**$15,699**	**$16,470**	**$16,833**
less: Annual maintenance (AHS)		$444	$445	$446	$470	$495	$510	$524	$519	$513	$501	$489	$493	$496	$506
less: Property taxes (AHS)		$1,564	$1,591	$1,619	$1,773	$1,928	$2,069	$2,210	$2,122	$2,034	$1,992	$1,950	$1,963	$1,976	$2,018
less: Homeowners insurance (AHS)		$461	$478	$495	$532	$570	$581	$592	$580	$569	$566	$564	$568	$571	$583
= Net operating income		**$9,563**	**$9,938**	**$10,313**	**$10,516**	**$10,744**	**$11,134**	**$11,397**	**$11,765**	**$11,899**	**$11,961**	**$12,357**	**$12,675**	**$13,427**	**$13,726**
less: Capital improvements (AHS)		$2,543	$2,815	$3,087	$3,472	$3,856	$3,604	$3,353	$3,332	$3,311	$2,974	$2,637	$2,655	$2,672	$2,728
less: Mortgage payments		$7,658	$7,658	$7,658	$7,658	$7,658	$7,658	$7,658	$7,658	$7,658	$5,171	$5,171	$5,171	$5,171	$5,171
= Imputed cash flow (Net benefit)		**–$639**	**–$536**	**–$433**	**–$613**	**–$770**	**–$128**	**$386**	**$775**	**$930**	**$3,817**	**$4,549**	**$4,850**	**$5,583**	**$5,827**
plus: Value of tax deducation (if itemize)		$2,371	$2,175	$2,159	$2,177	$2,193	$2,204	$2,213	$2,156	$2,097	$1,513	$1,483	$1,468	$1,453	$1,444
Imputed cash flow with tax benefit		**$1,732**	**$1,639**	**$1,726**	**$1,563**	**$1,423**	**$2,075**	**$2,599**	**$2,931**	**$3,028**	**$5,330**	**$6,032**	**$6,318**	**$7,036**	**$7,271**
Financial cash flows															
Value of home	$134,200	$141,900	$153,200	$169,500	$188,200	$195,600	$191,700	$177,900	$166,900	$157,900	$151,600	$155,400	$165,200	$172,200	$181,600
Cash to purchase	–$28,987														
Net sale proceeds (each year)		$26,006	$37,996	$54,722	$73,771	$82,408	$80,639	$69,771	$61,621	$55,452	$49,524	$54,882	$65,885	$74,354	$85,127
Annualized financial return on equity															
Internal rate of return on equity			12.6%	22.0%	24.9%	21.9%	17.3%	12.3%	9.1%	6.9%	5.8%	7.0%	8.4%	9.2%	10.0%
Internal rate of return with tax benefits			20.0%	28.4%	30.6%	27.2%	22.6%	17.8%	14.7%	12.7%	11.5%	12.3%	13.3%	13.8%	14.3%
Apartment Index after-tax returns			*23.1%*	*18.8%*	*21.5%*	*11.5%*	*5.7%*	*7.5%*	*10.1%*	*10.1%*	*9.5%*	*8.1%*	*9.5%*	*9.6%*	*9.0%*
S&P 500 Index after-tax returns			*14.1%*	*10.1%*	*10.5%*	*9.0%*	*0.4%*	*2.8%*	*3.8%*	*3.3%*	*4.1%*	*5.9%*	*6.3%*	*5.7%*	*5.9%*
Bond index after-tax returns			*5.2%*	*3.9%*	*3.6%*	*4.2%*	*4.0%*	*3.7%*	*3.7%*	*3.8%*	*3.6%*	*3.2%*	*3.2%*	*2.9%*	*2.8%*
Apartment index returns			30.0%	24.7%	28.3%	15.1%	7.2%	10.2%	14.3%	14.3%	13.6%	11.6%	13.7%	13.9%	13.1%
S&P 500 returns			17.4%	12.4%	12.7%	10.8%	0.4%	3.4%	4.6%	4.0%	4.9%	7.0%	7.3%	6.7%	6.9%
Bond index returns			6.9%	5.3%	5.0%	5.9%	5.7%	5.3%	5.4%	5.7%	5.5%	4.8%	4.9%	4.6%	4.4%

Notes: It is assumed the home buyer pays a 28 percent marginal tax rate on ordinary income and 20 percent on capital gains. Mortgage rates and costs are based on average annual data from Freddie Mac; initial mortgage is an 80 percent loan-to-value 30-year fixed rate in 2002 (5.83 percent plus 0.6 percent in points and 1.4 percent in other closing costs); and mortgage balance is refinanced in 2012 with a 3.6 percent 30-year fixed rate loan (with 0.7 percent in points and 1.3 percent in other closing costs). "Value of tax deduction" is marginal tax rate multiplied by sum of mortgage interest, property taxes, and points paid on mortgage. "Imputed rental 'income'" and "Value of home" are reported by Zillow for the median home in the United States. "Value of home" is as of December in each year. Rental income is the sum of rents over the course of the year when the data is measured. "Cash to purchase" is the cost of the home minus the mortgage amount plus points and closing costs on the mortgage. Sale proceeds are the sale price of the home, less 7 percent expense (for broker and other sale costs) and payoff amount on the mortgage. All amounts are compounded annually. After-tax returns for apartments are computed by using annual dividends from Real Estate Investment Trusts (REITs), estimating the taxable portion of the dividend, and paying the tax using the 28 percent marginal tax rate on ordinary income in the year the dividend was paid. Remaining earnings were taxed at the capital gains tax rate of 28 percent in the year of sale. Apartment REITs are companies that own and operate apartments and generally receive similar tax treatment as an individual investor would for an equivalent investment in apartment buildings. The S&P 500 returns were assumed to be entirely capital gains and taxed in the year of sale, whereas the returns on bond fund were taxed annually using a blended rate with capital gains and ordinary income.

the year the property is sold. All cash flows in these rows are undiscounted and measured at the end of each year. We compute the internal rate of return on home equity for the homebuyer assuming a sale in each year, with and without the tax benefit from itemized deductions.

Of course, any judgments about financial returns must take opportunity cost into account: that is, what a household might expect to earn on an investment of comparable risk if it decided to rent instead of purchase a home. Here we provide three possible benchmarks: an index of publicly traded apartment real estate investment trusts (REITs), the S&P 500, and a representative bond fund. In the last rows of Table 4, we report before- and after-tax returns for the comparable investments.

A note about after-tax returns: While most political debate about tax benefits of homeownership focuses on the tax deductions for mortgage interest and property taxes, even more important for many homeowners is the "hidden" benefit from not having to pay taxes on the imputed rent and capital gains on the home. Conversely, returns from investments in stocks and bonds are taxable, and we need to subtract household taxes for an apples-to-apples comparison of the financial return from owning a home. When it comes to investing in an apartment index, owners of rental properties are taxed on income from properties (including rents and fees) after deducting property expenses, including repairs and maintenance, depreciation, interest payments, and residential property taxes. When a rental home is sold, the owner pays capital gains taxes. In contrast, owner-occupants do not pay taxes on a capital gain up to $500,000 ($250,000 for singles) from the sale of their home under most circumstances.

The largest takeaway from the calculations in the table is that owning a home appears to be generally financially advantageous relative to renting, regardless of whether a homebuyer itemizes deductions. A homebuyer in 2002 would have earned a higher rate of return on home equity than on bonds regardless of the holding period, and a higher return than on the S&P 500 with a three-year holding period or more, once taxes on the alternative investment are considered. Including the value of deductions, the homebuyer would have outperformed all the alternative investments in all years. By contrast, that same buyer who did not itemize would have underperformed the publicly traded apartment real estate investment trust index for a two-year holding period and for holding periods ending in 2010–2015, a time period when demand for rental units was very high.

There are also important caveats. This analysis has focused exclusively on the returns for a representative national property over a single time period and thus doesn't incorporate what individual homeowners might have received on a specific property or in other time periods. It measures realized, not expected, returns. Moreover, new tax legislation may change the value of the tax benefits. As is often noted in investment prospectuses, past performance is not a guarantee of future returns.

What is driving these results? The last 15 years may have been a tough time period to invest in equities relative to real estate, as falling real interest long-term rates had a positive impact on returns for long-lived assets like housing. The strong

tax advantages associated with housing investments also play a role.[6] Another factor benefitting returns to homeowners is use of leverage to purchase a home. We assume a buyer uses a 20 percent down payment, which was the median at the beginning of 2003 according to Goodman et al. (2017), although first-time homebuyers put down less (and the median down payment in 2017 has declined to 12 percent).

The assumed mortgage embeds much higher leverage than is utilized by the typical apartment real estate investment trust, which might have 50 percent debt, or the leverage of a typical S&P 500 company. However, homeowners are able to take advantage of low borrowing costs associated with mortgages that typically have an implicit or explicit government guarantee, which is less-expensive debt than is available to corporate borrowers. It should be noted that high corporate leverage (or purchasing stocks using a margin account) is in many ways riskier than buying a home with high leverage. Individual investors and companies face potentially severe financial consequences of operating with high debt, including margin calls and debt downgrades and covenants that severely affect the ability of a company to function when leverage rises on a mark-to-market basis. By contrast, facing large costs of foreclosure and the loss of credit, many underwater homeowners were able to continue to make mortgage payments and wait for the housing market to recover. Indeed, as long as mortgage payments and other costs of owning a home are below the cost of renting an equivalent unit, an underwater homeowner has little financial incentive to default on a mortgage unless that homeowner would otherwise choose to downsize substantially.

Returns to Homeownership for Selected Metropolitan Areas

In Table 5, we calculate returns for owning in a few selected metropolitan areas. One limitation we face is that our analysis requires data from the American Home Survey in 2002–2004, which does not cover interesting housing markets such as Boston, Las Vegas, Miami, and San Francisco. Nonetheless, we are able to include data on Chicago, Cleveland, Dallas, Denver, Los Angeles, New York, and Phoenix. The analysis of the returns for these individual markets mostly mirrors the national data, with a few exceptions.

First, the average metropolitan area in these examples had higher average home prices and rates of home price appreciation than the United States as a whole, but lower average returns. In part, this finding suggests that a key component in understanding returns for purchasing a home comes from the rent/price ratio, which can be viewed as the initial cash yield on investment. Eisfeldt and Demers (2015) show that higher-priced homes have a lower cash yield on investment. Our analysis demonstrates that investing in a market with a high expected rate of appreciation may not have a strong financial return if the initial rental yield is sufficiently low. Also, commercial real estate investors in fast-growing markets often perceive these markets as having lower risk than the average market, as evidenced by low capitalization rates in so-called "gateway" (coastal) markets.

[6] Some individuals might choose to invest in tax-preferred vehicles like IRA or 401k. In this case, earnings are still taxed when the investor sells in retirement, but the effective tax rate would be lower than we estimate in this paper.

Table 5

Financial Returns from Owning a Home: Various Cities and National Data

Year	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*Annual % Δ 2003–16*
Dallas																
Value of home	$122,800	$127,200	$129,600	$139,000	$142,500	$146,400	$149,400	$138,200	$134,700	$137,400	$135,000	$139,300	$149,400	$159,100	$180,700	2.8%
IRR on equity with tax benefit			–3.3%	9.8%	11.1%	12.0%	12.2%	8.2%	7.4%	8.2%	7.7%	8.8%	10.2%	11.2%	12.5%	
Los Angeles																
Value of home	$271,900	$326,000	$393,500	$500,200	$587,400	$599,000	$552,800	$445,300	$420,100	$417,600	$388,700	$418,200	$491,500	$522,700	$557,600	5.3%
IRR on equity with tax benefit			60.7%	62.4%	54.3%	42.6%	31.8%	20.6%	16.8%	15.2%	12.6%	13.4%	15.1%	15.1%	15.1%	
Phoenix																
Value of home	$139,900	$148,400	$157,300	$187,300	$269,300	$266,200	$241,900	$196,100	$161,400	$139,800	$131,100	$161,300	$191,200	$198,600	$215,600	3.1%
IRR on equity with tax benefit			12.5%	31.3%	48.7%	37.3%	27.4%	16.9%	9.2%	4.1%	2.3%	8.4%	11.3%	11.5%	12.1%	
Cleveland																
Value of home	$128,100	$131,400	$137,700	$141,700	$144,800	$142,500	$135,600	$129,200	$125,500	$119,400	$113,700	$115,300	$117,700	$120,500	$124,800	–0.2%
IRR on equity with tax benefit			–4.6%	1.6%	4.0%	2.7%	–0.3%	–2.3%	–2.7%	–4.1%	–5.5%	–3.1%	–1.0%	0.6%	2.1%	
Denver																
Value of home	$214,200	$220,700	$223,500	$228,800	$234,900	$234,900	$230,900	$220,600	$217,400	$218,600	$214,600	$232,400	$253,500	$282,900	$324,200	3.0%
IRR on equity with tax benefit			–11.8%	–4.0%	0.1%	0.2%	–0.7%	–2.9%	–2.9%	–1.6%	–1.7%	2.1%	4.7%	7.1%	9.0%	
New York																
Value of home	$261,800	$292,100	$338,700	$386,700	$432,500	$443,700	$429,800	$393,200	$370,000	$357,300	$341,400	$341,700	$357,300	$364,900	$379,300	2.7%
IRR on equity with tax benefit			41.3%	42.3%	39.2%	32.5%	26.1%	20.0%	16.5%	14.5%	12.8%	12.5%	12.9%	12.9%	13.1%	
Chicago																
Value of home	$181,500	$195,700	$213,900	$224,700	$238,500	$246,400	$240,900	$228,500	$204,300	$186,500	$168,300	$166,600	$179,100	$186,900	$194,300	0.5%
IRR on equity with tax benefit			17.2%	17.9%	18.6%	17.1%	13.6%	10.1%	5.7%	2.9%	0.3%	1.8%	4.7%	6.2%	7.3%	
United States																
Value of home	$134,200	$141,900	$153,200	$169,500	$188,200	$195,600	$191,700	$177,900	$166,900	$157,900	$151,600	$155,400	$165,200	$172,200	$181,600	1.9%
IRR on equity with tax benefit			20.0%	28.4%	30.6%	27.2%	22.6%	17.8%	14.7%	12.7%	11.5%	12.3%	13.3%	13.8%	14.3%	

Notes: Data for individual cities as described in the text and calculated in Figure 4. IRR is internal rate of return. Value of home is reported by Zillow for the median home in the metropolitan area specified for the United States. All amounts are compounded annually. Marginal tax rates are assumed to be 28 percent in all Metropolitan areas, although they are likely higher in more coastal markets like New York and Los Angeles, and lower in Midwestern and Southern markets.

Returns to owning also depend critically on how much home price appreciation actually occurs. The slowest-growing markets like Cleveland and Chicago also had the lowest rate returns to owning a home, although only in Cleveland did the returns fall below the returns of the S&P 500. Of course, while the eventual realized relative returns were negative, it is unlikely that purchasers knew in advance which markets would rise or fall.

Finally, in all these markets, had a homeowner purchased in 2007, the returns would have been much lower than comparable stock market returns. Unless homebuyers can time the market (and choose the "right" city) with some foresight, purchasing a home is certainly not a guarantee of higher returns than renting. Academic papers such as Case and Shiller (1989) and Cochrane (2011) suggest there is a predictable component for returns to housing, although to some extent this predictability might be explained by time-varying risk preferences.

What Additional Risk and Benefits are Missing from These Financial Computations?

Along with the financial outcome, buying a home poses a range of other risks and benefits. Here, we discuss a number of issues associated with owning a home not included in the basic financial analysis: financial risks due to the concentration of wealth in a single asset; lock-in and decreased mobility effects; and homeownership as a method for disciplined savings and wealth accumulation. In fact, home equity is the principal source of savings for most American households, especially households in the bottom part of the income distribution, and ownership can serve to protect households from the financial risk of rising rents.

Of course, other factors might contribute to a high homeownership rate, but are missing from our discussion. For example, moral hazard concerns favor homeownership, because renters are unlikely to maintain a property as well as its owner would. Similarly, we cannot measure the many types of uncertainty that might affect owning a home in specific markets or explicitly compute whether the measured return provides sufficient excess return to compensate for perceived and actual risks. As well, in the past, a renter likely could not find a home of comparable quality to what was available to buy. But in the last decade, with the growth in institutional ownership of rental properties, there has been a renewed focus on providing rental homes that families desire in suburban locations with higher-quality school districts. There may also be cultural benefits from owning, and homeowners may develop an emotional attachment to their property that seems less likely in a rental property.

Financial Risks

Homeowners face potentially large financial risks associated with owning a single, undiversified, indivisible, sometimes illiquid asset that often represents the

vast majority of their wealth.[7] Piazzesi and Schneider (2016) offer an exhaustive summary of the many risks (and benefits) associated with homeownership. Households lack the ability to hedge either individual or aggregate movements in home prices. They face high transaction costs associated with moving, buying and selling homes, and foreclosures. Thus, households need a way to manage the risk of homeownership. Having the financial ability to weather the storms of volatility in home prices can be viewed as a method of effectively hedging volatility over time. In fact, few homeowners seem to feel the need to hedge price fluctuations. There have been a number of attempts to launch home price futures contracts, most recently the S&P/CaseShiller Home price contracts traded on the Chicago Mercantile exchange at the national level and for 10 cities, but these contracts have never gained much liquidity. More recently, a number of companies have been formed to sell home price insurance or a portion of home price appreciation, with little evidence of success. One potentially more promising market innovation is the attempt to embed home price and unemployment insurance explicitly into mortgages.

Sinai and Souleles (2005) point out an essential tradeoff between owning and renting: while owning exposes a household to home price risk, renting creates exposure to changes in rents. They show that the longer the expected time in a home, the lower the risk of owning relative to renting. In fact, some German renters purchase homes in a nearby neighborhood to take advantage of tax subsidies that favor owning rental property, which suggests that hedging rent risk is an important consideration for some middle-class renters. In a similar vein, Li and Yao (2007) discuss how house price changes can have differential effects depending on the age of the household.

Households also face risk related to mortgage financing and interest rates. Campbell and Cocco (2003) suggest that homeowners are often better-off taking out adjustable-rate instead of fixed-rate mortgages, although this choice is relatively uncommon. The fact that homeowners have not chosen to hedge risks that many economists estimate to be material, at least so far, suggests that this area is ripe for future research.

Lock-in and Decreased Mobility

One potential negative result of homeownership is impaired labor market mobility, especially in a downturn (Oswald 1996). The evidence on this possibility is mixed, at best.

One strand of this research has looked at correlations between homeownership and various labor market outcomes. Results appear at most to be small, and it has been hard to establish definitive results, which is perhaps not unexpected given the difficulties of disentangling cause and effect between homeownership and expected mobility. For example, some research has found some limited evidence (after adjusting for endogeneity issues) that homeownership is correlated with unemployment (Green and Hendershot 2001; Coulson and Fisher 2002, 2009;

[7] Innovations like Airbnb that allow a homeowner to rent a portion of the home provide new options for mitigating the financial risk of owning.

Van Leuvensteijn and Koning 2004; Munch, Roshholm, and Sarver 2006, 2008). More recently, Blanchflower and Oswald (2013) use state-level data with a fixed-effects model, finding that increases in the homeownership rate are followed by higher unemployment at the state level, although with long lags (up to five years). They also show that areas with high homeownership rates had lower labor mobility, longer commute times, and lower rates of business formation. Green and Wang (2015) present more complex findings, demonstrating that although homeownership may be slightly correlated with higher unemployment, it is also associated with longer employment spells, greater interstate mobility, and a lower likelihood of being unemployed. The inconsistent findings at the individual level at a minimum suggest a complex relationship that economic models have not fully captured.

A perhaps more promising strand of this literature examines whether specific circumstances such as negative equity, property tax benefits from staying, loss aversion, or low mortgage rates impair mobility. For example, Ferreira, Gyourko, and Tracy (2010) find that negative equity reduced mobility by 30 percent, and that each $1,000 of additional mortgage or property tax costs reduces household mobility by 10 to 16 percent (for earlier evidence, see also Genesove and Mayer 1997). However, using the same data but a different methodology, Schulhofer-Wohl (2011) argues that negative equity does not reduce mobility. Donovan and Schnure (2011) also find evidence of a lock-in effect, but argue that this effect is almost entirely driven by a decline in within-county moves, which are less likely to relate to moves that involve taking a new job. In contrast, out-of-state moves are higher in counties with greater home price declines, suggesting that falling home prices may even boost labor market mobility. Aaronson and Davis (2011) examine the post-recession timeframe from 2008 to mid-2010, a period of rising negative equity, and find no effect on interstate mobility. Consistent with Aaronson and Davis (2011), Sinai and Souleles (2013) show that households move between cities with highly correlated home prices, suggesting the lock-in is less likely to be an impediment to moving between metropolitan areas. Loss aversion also leads to a lower likelihood of selling a home when home prices fall (Engelhardt 2003; Genesove and Mayer 2001).

An overall reading of the existing evidence suggests that while specific factors related to falling home prices can impair mobility, these factors do not appear to meaningfully impede job-initiated moves. Moreover, given the expanded rental market for single-family homes, a homeowner now has an improved option to rent out the old home, find a rental property in the new location, and to postpone a decision to sell.

Homeownership and Wealth Accumulation

Homeownership has historically served as an effective vehicle for accumulating wealth for many reasons. Homes have generally appreciated in price over time. Owners typically pay down mortgage principal each month with nominally fixed payments that decline in real terms, can earn "sweat equity" by making improvements in their home, and benefit from favorable tax treatment (Herbert and Belsky 2008). Numerous studies show that homeowners have more wealth and accumulate wealth

faster than non-homeowners, although these effects are less pronounced for minority borrowers. Of course, it is quite difficult to disentangle correlation from causality.

Home equity is the largest component of net worth (excluding pensions and Social Security) and is particularly important for minority borrowers (Poterba, Venti, and Wise 2011, 2012). Median wealth of all homeowners in 2013 was $195,500, including $80,000 of home equity (Joint Center for Housing Studies 2015). Median home equity for white families was $90,000, 40 percent of median wealth for this group of $231,100. For black and Hispanic families, median wealth is much lower ($79,900 and $90,250, respectively) and home equity is even more important, representing more than half of that total ($47,000 and $48,000, respectively). Renters have relatively little net worth ($5,400). Pre-crisis studies showed that while homeownership carries significant risks, homeownership in the long term has been associated with strong wealth accumulation (Belsky and Duda 2002; Haurin and Rosenthal 2004; Herbert and Belsky 2008), particularly for those borrowers who have the willingness and ability to maintain homeownership during market fluctuations.

Of course, it is not clear how or whether homeownership contributes causally to wealth accumulation. After all, a number of studies done before the housing crisis in 2008 found that purchasing a home does not guarantee increases in wealth. The exit rate from homeownership was large for first-time, low-income borrowers—40 to 50 percent were unable to sustain homeownership for five years, with divorce being a major factor (Reid 2004; Haurin and Rosenthal 2005). Even controlling for observable characteristics that predict default like credit scores, loan purpose, loan-to-value ratio, debt-to-income ratio, and property characteristics (Haughwout, Peach, and Tracy 2008; Mayer, Pence, and Sherlund 2009), minority borrowers have been more likely to become delinquent on their mortgage loans with negative effects (Van Order and Zorn 2002; Deng and Gabriel 2006; Firestone, Van Order, and Zorn 2007; Fout, Li, and Palim 2017). In addition, home prices at the lower end of the market are more volatile than homes with higher prices (Piazzessi and Schneider 2016), exacerbating the size of wealth effects (positive and negative) for lower-income and minority borrowers who have higher-than-average loan-to-value ratios. Suburban locations with a high minority share of residents may also have lower appreciation rates than locations with a higher share of non-Hispanic white residents (Anacker 2010).

Post-2008 studies reaffirm the generally positive association between homeownership and wealth accumulation. Grinstein-Weiss, Key, Guo, Yeo, and Holub (2013) and Freeman and Ratcliffe (2012) study the Community Advantage Program, a program for low- and moderate-income borrowers, and find that after adjusting for outliers, the net worth of the new homeowners had increased more from 2005–2008 and fell less through 2010 than a matched group of renters. Herbert, McCue, and Sanchez-Moyano (2014, 2016) compare owners and renters using data from the Panel Study of Income Dynamics between 1999 and 2013. They find that homeownership was associated with significant gains in household wealth, although the magnitude of the gain was much smaller after the recession than before. They also find that a higher share of Hispanic and low-income households failed to sustain homeownership, while black households had smaller gains in wealth than other groups, after controlling for income, demographics, and household composition.

Turner and Smith (2009) also provide evidence that minority and low-income households are less likely to sustain homeownership, using data from the Panel Study of Income Dynamics from 1970 to 2005.

Attempts to disentangle correlation and causality between homeownership and household wealth are difficult. Sodini, Van Nieuwerburgh, Vestman, and von Lilienfeld-Toal (2016) address this endogeneity using a quasi-experiment from Sweden in which some residents are able to purchase their apartments at below-market prices. The paper shows that these homeowners become wealthier by saving more, have a relatively low marginal propensity to consume out of their newfound housing wealth, and invest more in equities. The paper attributes these effects predominantly to homeownership, although wealth effects also play a role.

Conclusion

Policymakers have traditionally viewed an expansion of homeownership as an important public policy goal, and they implemented policies during the 1990s and early 2000s to encourage homeownership. To the extent that anyone believed that all households should be homeowners, the financial crisis provided a strong counterexample illustrating the risks associated with homeownership when millions lost their homes to foreclosure. However, we have argued that homeownership remains very beneficial for most families, offering both financial gains and a chance to build wealth, especially for those who expect to own their homes for a long enough period of time to overcome transaction costs and near-term cyclical volatility. Today, it can be more difficult for households to become homeowners, reflecting difficulties in obtaining a mortgage, incomes that have not kept pace with the increases in home prices, as well as a lack of entry-level inventory in most housing markets. The restricted inventory of housing—due in large part to zoning restrictions, building codes and other issues—adds significantly to the costs of building a home. The public policy challenge in the United States should be to break down barriers that limit those who would benefit from homeownership from accessing it, while not pushing people to become homeowners for whom it doesn't make sense or providing subsidies where not appropriate.

■ *The opinions, analysis, and conclusions of this paper are those of the authors. Monica Clodius and Chris Hayes provided excellent research assistance. The paper benefitted from the help of Michael Lea, Deborah Lucas, Stephanie Moulton, Tomasz Piskorski, Mark Zandi, and the editor, Enrico Moretti; the excellent feedback from the reviewers Mark Gertler, Gordon Hanson, and Timothy Taylor; and data and assistance provided by Zillow and Svenja Gudell and Skylar Olsen. The research was supported by the Housing Finance Policy Center at the Urban Institute and the Paul Milstein Center for Real Estate at Columbia Business School. Goodman is on the board of directors of MFA Financial and is an advisor to Amherst Capital Management. Mayer is CEO of Longbridge Financial, a reverse mortgage lender.*

References

Aaronson, Daniel, and Jonathan Davis. 2011. "How Much Has House Lock Affected Labor Mobility and the Unemployment Rate?" Chicago Fed Letter 290, September.

Aaronson, Daniel, and Bhashkar Mazumder. 2005. "Intergenerational Economic Mobility in the U.S., 1940 to 2000." FRB Chicago Working Paper no. 2005-12, December.

Allison, Melissa. 2015. "Student Loan Debt Has a Minor Impact on Homeownership … As Long as You Get at Least a Four-Year Degree." *Zillow Porchlight*, September 16. https://www.zillow.com/blog/student-debt-effect-homeownership-182547/ February.

Anacker, Katrin B. 2010. "Still Paying the Race Tax? Analyzing Property Values in Homogeneous and Mixed-Race Suburbs." *Journal of Urban Affairs* 32(1): 55–77.

Badarinza, Cristian, John Y. Campbell, and Tarun Ramadorai. 2016. "International Comparative Household Finance." NBER Working Paper 22066. (And forthcoming in *Annual Review of Economics*.)

Barker, David, and Eric Miller. 2009. "Homeownership and Child Welfare." *Real Estate Economics* 37(2): 279–303.

Bond, Shaun A., and Michael D. Erikson. 2017. "The Role of Parents on the Home Ownership Experience of Their Children: Evidence from the Health and Retirement Study." University of Cincinnati Lindner College of Business Research Paper no. 2017-001. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3044708.

Belsky, Eric S., and Mark Duda. 2002. "Asset Appreciation, Time of Purchase and Sale, and the Return to Low-Income Homeowners." Chap. 7 in *Low-Income Homeownership: Examining the Unexamined Goal*, edited by Nicolas P. Retsinas and Eric S. Belsky. Washington, DC: Brookings Institution Press.

Blanchflower, David G., and Andrew J. Oswald. 2013. "Does High Home-Ownership Impair the Labor Market?" NBER Working Paper 19079, May.

Bogdon, Amy S. 1996. "Homeowner Renovation and Repair: The Decision to Hire Someone Else to Do the Project." *Journal of Housing Economics* 5(4): 323–50.

Bourassa, Steven C., and Martin Hoesli. 2010. "Why Do the Swiss Rent?" *Journal of Real Estate Finance and Economics* 40(3): 286–309.

Bush, George W. 2003. "President Bush Signs American Dream Downpayment Act of 2003." Remarks by the President at Signing of the American Dream Downpayment Act. Press Release, December 16, 2003. Department of Housing and Urban Development, Washington, DC.

Butrica, Barbara A., and Stipica Mudrazija. 2017. "Homeownership, Social Insurance, and Old-Age Security in the United States and Europe." Prepared for the 19th, Annual Joint Meeting of the Retirement Research Consortium, August 3–4, 2017. http://crr.bc.edu/wp-content/uploads/2017/08/5a.-Stipica-Mudrazija.pdf.

Campbell, John, and João F. Cocco. 2003. "Household Risk Management and Optimal Mortgage Choice." *Quarterly Journal of Economics* 118(4): 1449–94.

Case, Karl E., and Robert J. Shiller. 1989. "The Efficiency of the Market for Single-Family Homes." *American Economic Review* 79(1): 125–37.

Charles, Kerwin Kofi, and Erik Hurst. 2002. "The Transition to Home Ownership and the Black–White Wealth Gap." *Review of Economics and Statistics* 84(2): 281–97.

Cochrane, John H. 2011. "Presidential Address: Discount Rates." *Journal of Finance* 66(4): 1047–1108.

Coulson, N. Edward, and Lynn M. Fisher. 2002. "Tenure Choice and Labour Market Outcomes." *Housing Studies* 17(1): 35–49.

Coulson, N. Edward, and Lynn M. Fisher. 2009. "Housing Tenure and Labor Market Impacts: The Search Goes On." *Journal of Urban Economics* 65(3): 252–64.

Cynamon, Barry Z., and Steven M. Fazzari. 2014. "Inequality, the Great Recession, and Slow Recovery." Institute for New Economic Thinking Working Paper 9, October 1, 2014. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2638030.

Deng, Yongheng, and Stuart A. Gabriel. 2006. "Risk-Based Pricing and the Enhancement of Mortgage Credit Availability among Underserved and Higher Credit-Risk Populations." *Journal of Money, Credit, and Banking* 38(6): 1431–60.

Dietz, Robert D., and Donald R. Haurin. 2003. "The Social and Private Micro-level Consequences of Homeownership." *Journal of Urban Economics* 54(3): 401–450.

Donovan, Coleen, and Calvin Schnure. 2011. "Locked in the House: Do Underwater Mortgages Reduce Labor Market Mobility?" Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1856073.

Eisfeldt, Andrea, and Andrew Demers. 2015. "Rental Yields and HPA: The Returns to Single Family Rentals." NBER Working Paper 21804.

Englehardt, Gary V. 2003. "Nominal Loss Aversion, Housing Equity Constraints, and Household Mobility: Evidence from the United States." *Journal of Urban Economics* 53(1): 171–95.

Fannie Mae. 2014. "Fannie Mae National Housing Survey: What Younger Renters Want and

the Constraints They See." Slide presentation, May. http://www.fanniemae.com/resources/file/research/housingsurvey/pdf/nhsmay2014presentation.pdf.

Ferreira, Fernando, Joseph Gyourko, and Joseph Tracy. 2010. "Housing Busts and Household Mobility." *Journal of Urban Economics* 68(1): 34–45.

Firestone, Simon, Robert Van Order, and Peter Zorn. 2007. "The Performance of Low-Income and Minority Mortgages." *Real Estate Economics* 35(4): 479–504.

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

Fout, Hamilton, Grace Li, and Mark Palim. 2017. "Credit Risk of Low Income Mortgages." Economic and Strategic Research, Fannie Mae, May.

Freddie Mac. 2017. "Profile of Today's Renter." Multifamily Rental Research, March. http://www.freddiemac.com/multifamily/pdf/consumer_omnibus_results.pdf.

Freeman, Allison, and Janneke Ratcliffe. 2012. "Setting the Record Straight on Affordable Homeownership." Center for Community Capital Working Paper, May.

Garrison, Trey. 2015. "Moody's Analytics: Single-Family Rental Growth Will Accelerate: Biggest Growth in Western, Southern Markets." *HousingWire*, July 1, https://www.housingwire.com/articles/34369-moodys-analytics-single-family-rental-growth-will-accelerate.

Genesove, David, and Christopher Mayer. 1997. "Equity and Time to Sale in the Real Estate Market." *American Economic Review* 87(3): 255–69.

Genesove, David, and Christopher Mayer. 2001. "Loss Aversion and Seller Behavior: Evidence from the Housing Market." *Quarterly Journal of Economics* 116(4): 1233–60.

Gerardi, Kristopher, Andreas Lihnert, Shane M. Sherlund, and Paul Willen. 2008. "Making Sense of the Subprime Crisis." Brookings Papers on Economic Activity, Fall, 69–159.

Gicheva, Dora, and Jeffrey Thompson. 2015. "The Effects of Student Loans on Long-Term Household Financial Stability." In *Student Loans and the Dynamics of Debt*, edited by Brad Hershbein and Kevin M. Hollenbeck, 287–316. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Goodman, Laurie, Alanna McCargo, Ellen Seidman, Jim Parrott, Sheryl Pardo, Todd M. Hill, Jun Zhu, Bing Bai, Karan Kaul, Maia Woluchem, Bhargavi Ganesh, and Alison Ricon. 2017. "Housing Finance at a Glance: A Monthly Chartbook, May 2017." Research Report, Urban Institute, May 24. http://edit.urban.org/research/publication/housing-finance-glance-monthly-chartbook-may-2017.

Goodman, Laurie S., Rolf Pendall, and Jun Zhu. 2015. *Headship and Homeownership: What Does the Future Hold?* Urban Institute, June. http://www.urban.org/sites/default/files/publication/53671/2000257-Headship-and-Homeownership-What-Does-the-Future-Hold.pdf.

Goodman, Laurie S., Jun Zhu, and Bing Bai. 2016. "Overly Tight Credit Killed 1.1 Mortgages in 2015." *Urban Wire*, Urban Institute, November 21. http://www.urban.org/urban-wire/overly-tight-credit-killed-11-million-mortgages-2015.

Gordon, Robert J. 2012. "Is U.S. Economic Growth Over? Faltering Innovation Confronts the Six Headwinds." NBER Working Paper 18315, August.

Government of the District of Columbia, Office of the Chief Financial Officer, Office of Research and Analysis. 2003–2016. Annual reports titled "Tax Rates and Tax Burdens in The District of Columbia: A Nationwide Comparison" for the years 2002 to 2015. (We used 2003–2005.) Available at https://cfo.dc.gov/page/tax-burdens-comparison.

Grinstein-Weiss, Michal, Clinton Key, Shenyang Guo, Yeong Hun Yeo, and Krista Holub. 2013. "Homeownership and Wealth among Low- and Moderate-Income Households." *Housing Policy Debate* 23(2): 259–79.

Green, Richard K., and Hendershot Patrick. 2001. "Home–Ownership and Unemployment in the US." *Urban Studies* 38(9): 1509–20.

Green, Richard K., and Bingbing Wang. 2015. "Housing Tenure and Unemployment." June 1, 2015. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2628242.

Haughwout, Andrew, Richard Peach, and Joseph Tracy. 2008. "Juvenile Delinquent Mortgages: Bad Credit or Bad Economy?" *Journal of Urban Economics* 64(2): 246–57.

Haurin, Donald R., Christopher E. Herbert, and Stuart S. Rosenthal. 2007. "Homeownership Gaps among Low-Income and Minority Borrowers." *Cityscape* 9(2): 5–51.

Haurin, Donald R., and Stephanie Moulton. 2017. "International Perspectives on Homeownership and Home Equity Extraction by Senior Households," June 5. Available at: https://ssrn.com/abstract=2985917.

Haurin, Donald R., and Stuart S. Rosenthal. 2004. "The Impact of House Price Appreciation on Portfolio Composition and Savings." Washington, DC: US Department of Housing and Urban Development, Office of Policy Development and Research.

Haurin, Donald R., and Stuart S. Rosenthal. 2005. *The Sustainability of Homeownership: Factors Affecting the Duration of Homeownership and Rental Spells.* Washington, DC: US Department of

Housing and Urban Development, Office of Policy Development and Research.

Haurin, Donald R., and Stuart S. Rosenthal. 2007. "The Influence of Household Formation on Homeownership Rates across Time and Race." *Real Estate Economics* 35(4): 411–50.

Herbert, Christopher E., and Eric S. Belsky. 2008. "The Homeownership Experience of Low-Income and Minority Households: A Review and Synthesis of the Literature." *Cityscape* 10(2): 5–60.

Herbert, Christopher E., Donald R. Haurin, Stuart S. Rosenthal, and Mark Duda. 2005. *Homeownership Gaps among Low-Income and Minority Borrowers and Neighborhoods*. Washington, DC: US Department of Housing and Urban Development, Office of Policy Development and Research.

Herbert, Christopher, Daniel McCue, and Rocio Sanchez-Moyano. 2014. "Is Homeownership Still an Effective Means of Building Wealth for Low-Income and Minority Households?" In *Homeownership Built to Last: Balancing Access, Affordability and Risk after the Housing Crisis*, edited by Eric S. Belsky, Christopher E. Herbert, and Jennifer H. Molinsky. Cambridge, MA: Brookings Institution Press with the Joint Center for Housing Studies

Herbert, Christopher E., Daniel T. McCue, and Rocio Sanchez-Moyano. 2016. "Update on Homeownership Wealth Trajectories through the Housing Boom and Bust." Joint Center for Housing Studies Working Paper, Harvard University, February 18.

Holupka, Scott, and Sandra J. Newman. 2012. "The Effects of Homeownership on Children's Outcomes: Real Effects of Self Selection?" *Real Estate Economics* 40(3): 566–602

Hope Now. 2017 and back issues. Hope Now industry data reports. Most recent: http://www.hopenow.com/industry-data/HopeNow.FullReport.Updated(August2017).pdf.

Joint Center for Housing Studies. 2015. Appendix to *State of the Nation's Housing*. Table W-2: "Median Household Net Worth, Home Equity, and Non-Housing Wealth for Owners and Renters by Age and Race: 2013." Available at: http://www.jchs.harvard.edu/research/publications/state-nations-housing-2015.

Li, Wei, and Laurie Goodman. 2014. "Measuring Mortgage Credit Availability Using Ex-Ante Probability of Default." Research Report, Urban Institute, November 18.

Li, Wenli, and Rui Yao. 2007. "The Life-Cycle Effects of House Price Changes." *Journal of Money, Credit and Banking* 39(6): 1375–1409.

Lincoln Institute of Land Policy and the Minnesota Center for Fiscal Excellency. 2010–2014. *50-State Property Tax Comparison Study* for the years 2009–2013. Available at: http://datatoolkits.lincolninst.edu/subcenters/significant-features-property-tax/resources.aspx.

Lincoln Institute of Land Policy and the Minnesota Center for Fiscal Excellency. 2015–2017. *50-State Property Tax Comparison Study* for the years 2014 to 2016. Available at: http://www.lincolninst.edu/search/site/50-State%2520Property%2520Tax%2520Comparison%2520Study?f%5B0%5D=mediatype_levelone%3A5.

Lu, Chenxi, and Eric Toder. 2016. "Effects of Reforms on the Home Mortgage Interest Deduction by Income Group and State." Tax Policy Center, Urban Institute and Brookings Institute, December 6. http://www.taxpolicycenter.org/sites/default/files/publication/136906/2001015-effects-of-reforms-of-the-home-mortgage-interest-deduction-by-income-group-and-by-state_0.pdf.

Mayer, Christopher. 2017. "Housing, Mortgages, and Retirement." Chap. 9. in *Evidence and Innovation in Housing Law and Policy*, edited by Lee Anne Fennell and Benjamin J. Keys. Cambridge University Press.

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

McKernan, Signe-Mary, Caroline Ratcliffe, C. Eugene Steuerle, and Sisi Zhang. 2013. "Less than Equal: Racial Disparities in Wealth Accumulation." Urban Institute, April. http://www.urban.org/sites/default/files/publication/23536/412802-less-than-equal-racial-disparities-in-wealth-accumulation.pdf.

Mian, Atif, and Amir Sufi. 2009. "The Consequences of Mortgage Credit Expansion: Evidence from the US Mortgage Default Crisis." *Quarterly Journal of Economics* 124(4): 1449–96.

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

Minnesota Center for Fiscal Excellence. 2007–2009. *50-State Property Tax Comparison Study* for years 2006–2008. Available at: https://www.fiscalexcellence.org/our-studies.html;jsessionid=60B56CA25AECEDA80F7AB396CD2A585A?page=3.

Munch, Jakob Roland, Michael Roshholm, and Michael Svarer. 2006. "Are Homeowners Really More Unemployed?" *Economic Journal* 116(514): 991–1013.

Munch, Jakob Roland, Michael Roshholm, and Michael Svarer. 2008. "Home Ownership, Job Duration, and Wages." *Journal of Urban Economics* 63(1): 130–145.

National Association of Realtors Research Department. 2017. "Aspiring Home Buyers Profile." February. https://www.nar.realtor/reports/aspiring-home-buyers-profile.

National Association of Real Estate Investment Trusts (Nareit). 2017. *Monthly Property Index Values & Returns, Apartments*. Available from: https://www.reit.com/data-research/reit-indexes/

monthly-property-index-values-returns.

Nothaft, Frank E., and Yan Chang. 2004. "Refinance and the Accumulation of Home Equity Wealth." Freddie Mac Working Paper no. 04-02, February.

Oswald, Andrew J. 1996. "A Conjecture on the Explanation for High Unemployment in the Industrialized Nations: Part I." Warwick Economic Research Papers 475, University of Warwick.

Phillips, Matt. 2014. "Most Germans Don't Buy Their Homes, They Rent. Here's Why." *Quartz*, January 23. https://qz.com/167887/germany-has-one-of-the-worlds-lowest-homeownership-rates/.

Piazzesi, Monika, and Martin Schneider. 2016. "Housing and Macreconomics." Stanford University Working paper, July. http://web.stanford.edu/~piazzesi/housingandmacroeconomics.pdf. (And forthcoming in *Handbook of Macroeconomics.*)

Poterba, James, Steven Venti, and David Wise. 2011. "The Composition and Drawdown of Wealth in Retirement." *Journal of Economic Perspectives* 25(4): 95–118.

Poterba, James M., Steven F. Venti, and David A. Wise. 2012. "Were They Prepared for Retirement? Financial Status at Advanced Ages in the HRS and AHEAD Cohorts." In *Investigations in the Economics of Aging*, edited by David A. Wise, p. 21–69. University of Chicago Press.

Reid, Carolina. 2004. "Achieving the American Dream? A Longitudinal Analysis of the Homeownership Experiences of Low-Income Households." CSDE Working Paper no. 04-04, University of Washington, Seattle.

Schulhofer-Wohl, Sam. 2011. "Negative Equity Does Not Reduce Homeowners Mobility." NBER Working Paper 16701, January.

Schwartz, Heather L., Raphael W. Bostic, Richard K. Green, Vincent J. Reina, Lois M. Davis, Catherine H. Augustine. 2016. "Preservation of Affordable Rental Housing: Evaluation of the McArthur Foundation's Window of Opportunity Initiative." RAND.

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

Sinai, Todd, and Nicholas Souleles. 2013. "Can Owning a Home Hedge the Risk of Moving?" *American Economic Journal: Economic Policy* 5(2): 282–312.

Simmons, Patrick. 2014. "Upper-Income, Educated, Married with Children, and Still Not Buying: Declining Homeownership among 'Prime' First-Time Home Buying Candidates." Fannie Mae Housing Insights Brief. *Fannie Mae Economic and Strategic Research* 4(4), August 18, 2014. http://www.fanniemae.com/resources/file/research/datanotes/pdf/housing-insights-081814.pdf.

Sodini, Paolo, Stijn Van Nieuwerburgh, Roine Vestman, and Ulf von Lilienfeld-Toal. 2016. "Identifying the Benefits from Home Ownership: A Swedish Experiment." NBER Working Paper 22882.

Turner, Tracy M., and Marc T. Smith. 2009. "Exits from Homeownership: The Effects of Race, Ethnicity, and Income." *Journal of Regional Science* 49(1): 1–32.

Urban Institute. 2015. "Nine Charts about Wealth Inequality in America." Urban Institute. http://apps.urban.org/features/wealth-inequality-charts/.

US Department of Housing and Urban Development. 1995. *The National Homeownership Strategy: Partners in the American Dream*, Washington DC, May.

US Census Bureau. 2016a. "American Housing Survey (AHS): About." Webpage, October 25. https://www.census.gov/programs-surveys/ahs/about.html.

US Census Bureau. Various years. *AHS National and Metropolitan Public Use File (PUF),* CSV file for years 2002 to 2015. Available at: https://www.census.gov/programs-surveys/ahs/data.All.html.

Van Leuvensteijn, Michiel, and Pierre Koning. 2004. "The Effect of Home-Ownership on Labor Mobility in the Netherlands." *Journal of Urban Economics* 55(3): 580–96.

Van Order, Robert, and Peter Zorn. 2002. "Performance of Low-Income and Minority Mortgages." Chap. 11 in *Low-Income Homeownership: Examining the Unexamined Goal*, edited by Nicolas P. Retsinas and Eric S. Belsky. Brookings Institution Press: Washington, D.C.

Yahoo! Finance. 2017a. *iShares Core US Aggregate Bond (AGG): Monthly Historical Prices,* CSV file. Available at https://finance.yahoo.com/quote/AGG/history?p=AGG.

Yahoo! Finance. 2017b. *S&P 500 (^GSPC): Monthly Historical Prices,* CSV file. Available at: https://finance.yahoo.com/quote/%5EGSPC/history/.

Zillow, Inc. 2016. *Quarterly Historic Metro ZRI,* CSV file. Available at: https://www.zillow.com/research/data/.

Zillow, Inc. 2017. *ZHVI Single-Family Homes Time Series ($): METRO/US,* CSV file. Available at: https://www.zillow.com/research/data/.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 59–80

Sand Castles Before the Tide? Affordable Housing in Expensive Cities

Gabriel Metcalf

For decades following World War II, America's urban crisis was one of decline and population loss—problems that persist in some US cities. But the 1990 Census showed that an important set of cities had begun to gain population over the previous decade (and some neighborhoods had begun to attract new residents even earlier). Crisis became renaissance, and these cities came to experience an entirely different set of problems.

Today, we observe the divergent fates of American cities: some are becoming extremely costly while others continue to struggle with the problems of abandonment; some grow at a rapid pace while others resist new development. Broadly speaking, we can classify US cities into three types in terms of their housing cost dynamics. First, some cities continue to have shrinking populations, so the existing supply of housing is large compared to the quantity demanded and housing is often quite inexpensive. Examples include certain "Rust Belt" cities like Rochester, Detroit, and St. Louis. Second, some cities have both growing population and a growing supply of housing, including "Sun Belt" cities such as Atlanta, Houston, and Tucson. These cities tend to have relatively less-expensive housing. Third, in some cities, the demand for housing is growing at a much faster rate than the supply. These so-called "superstars" include New York City, Boston, Washington, DC, San Francisco, Los Angeles, Seattle, and Denver (Gyourko, Mayer, and Sinai 2013). Table 1 shows housing price increases over the past 20 years for 17 large metro areas.

■ *Gabriel Metcalf is President of SPUR, an urban policy research and advocacy organization in the San Francisco Bay Area, California. The web address is http://www.spur.org.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.1.59

doi=10.1257/jep.32.1.59

Table 1
Change in Median Home Values 1996 to 2016

Core-based Statistical Area	*Median value in 1996 (2016 USD)*	*Median value in 2016 (2016 USD)*	*Percent increase*
San Francisco–Oakland–Hayward, CA	$302,926	$813,108	168%
Los Angeles–Long Beach–Anaheim, CA	$229,135	$576,200	151%
San Diego-Carlsbad, CA	$219,981	$515,325	134%
Riverside–San Bernardino–Ontario, CA	$150,947	$310,433	106%
Boston–Cambridge–Newton, MA–NH	$203,048	$399,100	97%
Seattle–Tacoma–Bellevue, WA	$204,289	$396,717	94%
Denver–Aurora–Lakewood, CO	$177,498	$341,292	92%
Miami–Fort Lauderdale–West Palm Beach, FL	$125,039	$236,867	89%
Washington–Arlington–Alexandria, DC–VA–MD–WV	$207,790	$372,375	79%
New York–Newark–Jersey City, NY–NJ–PA	$223,167	$390,275	75%
Tampa–St. Petersburg–Clearwater, FL	$ 99,863	$169,908	70%
Phoenix–Mesa–Scottsdale, AZ	$143,303	$223,392	56%
Minneapolis–St. Paul–Bloomington, MN–WI	$150,259	$229,117	52%
Philadelphia–Camden–Wilmington, PA–NJ–DE–MD	$142,929	$209,900	47%
Dallas–Fort Worth–Arlington, TX	$136,317	$192,150	41%
St. Louis, MO–IL	$110,619	$143,917	30%
Atlanta–Sandy Springs–Roswell, GA	$144,201	$167,467	16%

Sources and Notes: Median values are based on Zillow Median Housing Value Index for all homes by Core-based Statistical Area. Monthly medians were collapsed to annual average medians. Figures were deflated to 2016 USD using CPI-U from the Bureau of Labor Statistics. List of Core-based Statistical Areas is based on the 20 largest Metropolitan Statistical Areas by population in 2016 (American Community Survey 2016 1-year estimates, Table B01003). Chicago, Detroit, and Houston CBSAs are not listed due to lack of data.

This third set of cities, with unprecedented economic success and a seemingly permanent crisis of affordable housing, is the focus of my article. In the expensive cities, policymakers expend great amounts of energy trying to bring down housing costs with subsidies for affordable housing and sometimes with rent control. But these efforts are undermined by planning decisions that make housing for most people vastly more expensive than it has to be by restricting the supply of new units even in the face of growing demand.

I begin by describing current housing policy in the expensive metro areas of the United States. I then show how this combination of policies affecting housing, despite internal contradictions, makes sense from the perspective of the political coalitions that can form in a setting of fragmented local jurisdictions, local control over land use policies, and homeowner control over local government. Finally, I propose some more effective approaches to housing policy.[1]

[1] How do we know if housing is "too expensive?" There are three common ways that practitioners think about affordability. First, the US Department of Housing and Urban Development defines a household as "cost-burdened" if it pays more than 30 percent of its pretax income for housing. Second, the Center for Neighborhood Technology (CNT) proposes a measure that combines household expenses on housing

We are interested in both metro areas and in the individual cities that make up metro areas. Housing markets and labor markets—conceptually, the same thing in most cases—exist at the scale of metropolitan regions. Because people within a metropolitan area can easily live in one city but work in another, it's not possible to bring down the cost of housing in one city without bringing it down in the metro region as a whole. But as we will see, the decisions that affect housing costs are not made at the metropolitan scale, they are made at the scale of individual cities. So it is usually correct to speak about the housing policy choices of *cities*, even when the outcomes of those policy choices will be manifest at the metropolitan scale. I will try to be clear about scale throughout this discussion.

Overall, my view is that the effects of the formal affordable housing policies of expensive cities are quite small in their impact when compared to the size of the problem—like sand castles before the tide. I will argue that we can do more, potentially much more, to create subsidized affordable housing in high-cost American cities. But more fundamentally, we will need to rethink the broader set of exclusionary land use policies that are the primary reason that housing in these cities has become so expensive. We cannot solve the problem unless we fix the housing market itself.

Urban Housing Policy Today

Cities have four principal tools they use to affect housing prices: direct provision of social housing; vouchers to increase the purchasing power of households; price controls on rents; and regulations on development of new housing supply. We'll review each of these.

Social Housing

In the early days of the affordable housing movement in America, many activists argued that housing for the broad working and middle class should be provided outside the market. They drew favorable lessons from European cities such as Amsterdam and Vienna. Activists like Catherine Bauer (1934, p. xvi) wrote approvingly of the European models, saying, "The land, construction, finance, and management of low- and medium-cost dwellings were removed from the speculative market: housing became a utility."

---

and transportation, as a percent of household income (Haas, Makarewicz, Benedict, Sanchez, and Dawkins 2006). CNT's work shows that metros with lower housing costs tend to have higher transportation costs; and metros that are more expensive on the "H+T" (Housing and Transportation Affordability) Index also have higher average household incomes—because housing costs partially determine wage rates necessary to attract workers. Third, Stone, Burke, and Ralston (2011) has proposed a "residual income" approach that defines housing affordability by how much money a household has left over after all nondiscretionary expenses, including housing. In this essay I take a qualitative view. I suggest that when we say that housing is "too expensive," what we mean is that we want more people to be able to live in the high-productivity metro areas of America; in other words, we want people lower down on the skill ladder (or people without as much inherited wealth), to have the chance to be part of these successful agglomeration economies.

Table 2
Proportion of Housing Units Receiving Subsidies or under Rent Regulation for Selected US Cities

	Social Housing Sector				
	Public housing	*Other subsidized*	*Rent regulated*	*Unregulated rental*	*Owner occupied*
New York City	6.0%	4.0%	34.6%	25.1%	30.2%
Los Angeles	1.7%	3.4%	14.0%	42.3%	38.6%
Chicago	3.8%	2.6%	0.0%	49.8%	43.8%
Houston	1.0%	3.4%	0.0%	49.8%	45.8%
Philadelphia	3.7%	1.5%	0.0%	35.5%	59.3%
Phoenix	1.0%	1.9%	0.0%	36.5%	60.7%
San Diego	1.4%	3.4%	0.2%	45.5%	49.5%
Dallas	0.5%	2.9%	0.0%	53.3%	43.2%
San Antonio	2.8%	2.1%	0.0%	37.0%	58.1%
Detroit	2.0%	3.9%	0.0%	39.2%	54.9%
San Francisco	1.2%	6.8%	46%	9%	36%

Source: Based on Ellen and O'Flaherty (2013, Tables 10.1 and 10.4), who use data from the American Housing Survey Metropolitan Data series of the US Census Bureau. San Francisco data is from the San Francisco Mayor's Office of Housing by communication with the author.

The contrast between US housing policy and that in Western Europe remains instructive: today we see high levels of social housing in the Netherlands (33 percent), France (17 percent), Denmark (20 percent), and the United Kingdom (18 percent) (Housing Europe 2015). In these countries, it is much more common for working-class and middle-class people (not just the very poor) to live in social housing. The sector contains a mix of publicly owned housing, publicly funded (but privately owned) housing, and cooperative housing, with lots of specific differences in institutional design across countries.

Table 2 loosely translates this idea into an American context with data for ten US cities. I define the "social housing" sector as housing that is both subsidized and permanently price-restricted.

In the United States, following some scattered experiments during World War I, large-scale construction of public housing began in earnest with the 1937 Wagner Housing Act, which launched both public housing and urban renewal as part of a national effort to tear down "slum" housing and replace it with "modern" housing (Radford 1996). The heyday of the program, in terms of how many units were produced, was the 1950s and 1960s. But by the end of the 1960s, public housing was in disrepute—the result of bad design (public housing became a playground for architectural fads); racism (one of the reasons Congress never adequately funded the program); public sector sclerosis (poor management by local housing authorities); broader economic decline in these local areas (which left residents in deeper poverty over time); and perhaps an underlying faulty premise about the efficacy of concentrating so many poor people in one location.

Many reforms of public housing were launched over the years, perhaps the most extensive being the HOPE VI program of 1993–1999, which offered block grants to cities to replace the modernist "towers in the park" with low-rise, more traditional buildings like row houses (Solomon 2003).

Addressing the mistakes of previous generations of public housing continues to be a major focus of housing policy in most cities. In the late 1960s and early 1970s, most cities began to contract out the construction and management of their subsidized housing programs. The Community Development Corporation (CDC) emerged as the primary organizational model for this work, bringing greater control by local community leaders, linking housing to a broader agenda of neighborhood revitalization, and introducing better management practices that marked a significant advance over the public housing authorities (Erickson 2009).

In more recent years, some cities (including New York, Washington, DC, Boston, Portland, Los Angeles, and San Francisco) have provided social housing through a policy of "inclusionary housing," which requires that market-rate housing developers set aside a portion of their units at below-market prices permanently or pay an equivalent fee (Joseph, Chaskin, and Webber 2007). Typically, these laws require between 5 and 25 percent of the units in a market-rate building to be provided at below-market rents (activists always hope for more, and sometimes get it). The cost of each inclusionary unit can be quite high, ranging between $250,000 per inclusionary unit in a lower-cost building to more like $500,000–$700,000 dollars in subsidy per inclusionary unit in a new high-rise. When a developer pays a fee rather than build the inclusionary units on site, the costs can also be significant; in San Francisco in 2016 the fee for each two-bedroom inclusionary unit built off-site was $366,000. On a 100-unit building, with a 15 percent inclusionary requirement, the total fee would be $5,490,000 (San Francisco Office of the Controller 2016). Those fees have continued to rise.

The inclusionary units are allocated by a lottery, with hundreds or even thousands of people applying for each one. Inclusionary housing has been important in a few markets, but the numbers of units that can be generated through this type of program are exceedingly small, because: a) the internal economics of the developments can support only so many below-market-rate units; and b) the taxable base—the number of market-rate projects that are built in any year—is not usually very large. Every now and then people suggest that more households could be helped if the subsidies from inclusionary programs were spent in less-expensive municipalities within the same metro region, as when the Mayor of Oakland suggested that San Francisco's affordable housing dollars would make a bigger difference if they were spent in Oakland. Thus far, policymakers have decided that it is more important to spend these resources to further the goal of income diversity within their cities, in keeping with the concept of "inclusion." Useful reports that compile data on US inclusionary housing programs include Hickey, Sturtevant, and Thaden (2014), Sturtevant (2016), Jacobus (2015), and Williams, Carlton, Juntunen, Picha, and Wilkerson (2016).

In theory, social housing has an important societal benefit beyond the improved well-being of the residents who get to live in it: by placing urban land into ownership by nonmarket actors, it provides one way to address issues related to the economic rents of landholding. The gains in housing wealth generated by the highly productive cities since the "great divergence" of the 1970s were not a reward for hard work or innovation on the part of landowners. The efforts of homeowners to use regulatory tools to protect and extend these wealth gains are the epitome of wasteful rent-seeking behavior.[2] Public and nonprofit ownership of urban land provides a direct way (although certainly not the only way) to remove some parcels from the rent-seeking behavior by private landowners in a context of housing scarcity. (For a modern discussion of Henry George's famous proposal to tax away the "un-earned" increment of land value, see Arnott and Stiglitz 1979.)

Social housing could be expanded in novel ways, including serving a broader range of income levels as in the European models. In Metcalf (2015), I offer a history of the concept of "alternative institutions" in American social movements, various attempts to invent new models of affordable housing. The natural question with this approach is one of funding, and the numbers can become forbiddingly large. The math is conceptually simple: multiply the subsidy per unit by the number of units we want to build or acquire. Assuming a subsidy of $300,000 per unit (it can be much more in high-cost cities), if we want to help one million households the cost would be $300 billion.

Vouchers

In 2015, 2.2 million households, comprising 5 million people, used rental vouchers to secure housing in the private market. The biggest program known as "Section 8," was created in 1974. Under the Section 8 program, households pay 30 percent of their income in rent, and the local Housing Authority covers the rest of the monthly rent to the landlord. Each year, the US Department of Housing and Urban Development determines "fair market rent" which sets the limits on how much rent subsidy will be provided in each city. (As of this writing in 2017, the HUD fair market rent for a two-bedroom unit in San Francisco is $3,319 per month.)

The federal government does not fund vouchers for everyone who needs them, and there are long waiting lists in most cities. One study estimates that only 25 percent of the households that are income eligible according to the standards of the US Department of Housing and Urban Development receive federal assistance (Center on Budget and Policy Priorities 2017). In some cities, the odds are much worse: recently, 600,000 residents of Los Angeles were applying for 2,400 vouchers (Smith 2017). In expensive housing markets, there is also a perennial problem of

[2]Land rents can be roughly estimated as the difference between housing sales prices and the full cost of production. In an unconstrained housing market, prices should decline to approach marginal costs (Gyourko and Molloy 2014). Rognlie (2015) shows how the changing distribution of wealth in the form of land rents to the housing sector is a major cause of the changes in inequality of wealth since World War II (see also Avent 2015).

voucher dollar amounts being insufficient, so that many landlords are not willing to rent to voucher holders.

Cities generally do not invest their own affordable housing funds into expanded voucher programs, preferring to instead create permanent social housing units, but a local expansion of housing voucher programs remains an option that could be pursued.[3] Indeed, this program could be expanded by any level of government (local, state, or national) if the political will existed.

In theory, vouchers have many virtues. They allow targeting of benefits to the people who most need them. When provided at the federal or state level, they can be used in many different locations, opening up different neighborhoods and school districts to people from different economic backgrounds. They are flexible in the depth of subsidy they provide based on the exact income of each household.

In practice, the program does not work as well as we might wish. There is pervasive discrimination against voucher holders, such that in many places certain landlords specialize in housing the population of voucher holders. In low-elasticity housing markets, vouchers can end up increasing the cost of housing, whereas direct provision of social housing can expand the supply and can drive down prices in the lower end of the housing market. To truly reach its potential as a tool for lifting low-income families out of poverty by giving them access to better school districts and other opportunities, voucher programs need to be supported by more intensive counseling programs and other forms of assistance (DeLuca and Rosenblatt 2017).

Given the costs of both social housing and vouchers, there is reason to wonder if the United States would ever have the political will to spend enough money on housing subsidies to help everyone who needs it, especially when we consider the question of trade-offs: Are we certain that it's best to spend that money on housing as opposed to say, education? But remember that the federal government spends far more on subsidies for homeowners than it does on subsidies for renters, this in the form of the mortgage interest deduction ($71 billion), the deduction for real estate taxes ($31.4 billion), and the tax exclusion on capital gains from housing ($24.1 billion). Taken together, these numbers from 2015 totaled more than double the combined costs of support for low-income non-homeowners like Section 8 housing vouchers ($29 billion), the low-income housing tax credit ($7.6 billion), public housing ($6.5 billion), and accelerated depreciation ($4.7 billion), which is a tax benefit for rental apartment owners who use federal low-income tax credits (Fischer and Sard 2016; Schwartz 2015). We can surely spend more money on vouchers and/or social housing if we choose to.

We are already making an investment in housing subsidies at a massive scale. Perhaps it is no more unreasonable to hope for a truly large social housing program or voucher program than it is to wish for a truly large change to the local rules on housing supply; both are uphill fights.

[3]For a classic review of the debate between subsidizing the production of housing or subsidizing the purchasing power of households, see Apgar (1990). For interesting thinking about how to redesign the housing voucher program, see Collinson and Ganong (2017).

Rent Control

Rent control is relatively rare in American cities and occurs mainly in the states of New York, New Jersey, and California. In our cohort of expensive cities, rent control is especially significant in New York, San Francisco, and Los Angeles. From the sample of economists that I have known, it appears that opposition to rent control is something like an oath of office for the profession, but real-world rent control, at least in its modern form, is generally not very damaging in its impacts on the housing market (Arnott 1995). Generally, landlords are allowed to raise the rent a certain percent each year for existing tenants, and there are rules to prevent landlords from evicting tenants without "just cause." But landlords can usually raise the rents up to market rate, with no restrictions, upon unit vacancy. Nowhere in the United States does rent control apply to new construction. In a sense, rent control works as a delay mechanism that slows the rate of price increases on incumbent tenants for part of the housing stock. This American version of rent control is quite different from rent control in places like Paris, where the government sets the allowable maximum rent each year for all the regulated units (O'Sullivan 2016).

We should acknowledge the downsides of US-style rent control. It limits unit turnover and leads to a misallocation of housing resources. It has poor targeting efficiency in terms of matching the benefits to the people who most need them. It adds to the perception of risk (and the cost of capital) for investors in new development, who will fear that cities with pro-rent control politics could at some point try to apply it to new construction or otherwise change the universe of units that fall under the price controls. It benefits current residents while doing nothing for new migrants to cities. But where rent control has been in place for a while, it is not typically a major cause of supply suppression. So long as cities are not trying to apply rent control to new (or recently built) development, it is a sidebar to the more fundamental dynamics that affect the cost of housing in expensive metro areas. Against these downsides, we should also acknowledge the significant upsides of large groups of people enjoying lower housing rents than they otherwise would, with the attendant benefits of greater community stability.

Regulation of the Housing Market

Given how much effort cities put into their official affordable housing programs, it is paradoxical, or even tragic, that when we turn to housing policy for the market-rate sector we find that the preponderance of the effort is geared toward suppressing supply. Local development regulations fall into four categories: zoning, building standards, permits to add supply, and fees.

Zoning codes regulate what land uses are allowed on a site—housing, office, retail, and so on. They also control building heights, densities (how much building per area of land is allowed), set-backs, rear-yard requirements, tower separation requirements, parking requirements, and other aspects of building use and form. Between historic districts, solar protection rules, and hundreds of other controls—a broader set of regulations than the mere designation of "zones"—the rules can become quite complex. Zoning in America is generally delegated to locally elected

legislative bodies, like city councils, although in many cities, especially in California, zoning ordinances can also be enacted by ballot initiatives (Fischel 2015).[4]

If zoning regulates *what* can built where, the building code (and other related codes) regulate *how* it can built: what materials are allowed, how big the windows must be, how large the rooms must be, how much heat can be lost through a wall, how a structure performs in an earthquake, and so on. Such technical regulations inevitably have both benefits and costs, which can be difficult to assess. Many of these codes are necessary, but they have the effect of raising the production costs of housing. Are the added costs worth it? In some cases, the answer will be no. Especially for those who hope that innovation will lead to reduced housing production costs, building standards will often prove to be a barrier.

Both zoning rules and building codes embody judgments regarding what constitutes "decent" housing. The rules inherently involve subjective criteria about aesthetics and livability. For example, many cities effectively outlaw single room occupant apartments, rooming houses, and other shared housing models that once provided cheap housing to the working class (Groth 1994). Those who believe that one strategy to bring down the costs of housing should be to allow people to live in smaller and less-expensive types of housing, may feel that the minimal standards have not been set in the right place.

The housing approval process is the next piece of the puzzle. A developer can propose to build housing that fits within the zoning code, the building code, and all the other codes, but must also still receive legal permission to build something. The process for getting this permission (or "entitlement") varies widely across cities, in what can be viewed as a continuum of certainty. Some jurisdictions allow housing that fits within the zoning codes to be approved automatically. In other places—again, California cities stand out—a developer proposing a large project will need to pay for years of studies about environmental impacts; hold dozens of public meetings at which neighbors express their desires for the project to be changed, reduced, or rejected; hire lobbyists, make campaign contributions, and donate money to community groups to convince elected officials to allow the project; and ultimately face a vote of the city council to allow or disallow the project. After that, in some jurisdictions, the project may still end up on the ballot to face a vote of the entire electorate. More uncertainty and greater risk translates into a higher cost of capital. Longer approval processes translate into higher carrying costs for the land.

[4]The 1926 US Supreme Court case that established the validity of zoning, *Village of Euclid, Ohio v. Ambler Realty Co.* (272 US 365), established the "presumption of validity" that locally-elected legislative bodies were to be treated as the judges of what was in the public interest. But the court also noted that there could come a time in the future when what might be perceived to be good for a municipality would diverge from the broader public interest: "It is not meant by this, however, to exclude the possibility of cases where the general public interest would so far outweigh the interest of the municipality that the municipality would not be allowed to stand in the way." This idea has reappeared in many important land use cases: as another example, see the 1972 case before the New York Court of Appeals, *Golden v. Planning Board of Town of Ramapo* (285 N.E. 2d). It remains to be seen whether housing reformers will be able to develop a legal strategy based on the insight that the broader regional or national public interest is not necessarily aligned with the incentives of individual cities.

Perhaps the greatest negative impact of an uncertain and hyperpoliticized entitlement process is that it functions as a barrier to entry for developers and investors into a market. The net effect is to reduce competition among developers.

The fourth type of local regulation on housing development is financial: fees and exactions. The legal distinctions between these types of payments are important for city officials and developers, but the economic logic is similar: these payments must be made in exchange for permission to build housing. Cities collect fees and exactions to support affordable housing production, transit expansion, parks, and general municipal budgets. The total costs of fees and exactions in a city like San Francisco range between $60,000 and $150,000 for each market-rate unit.

These costs interact with the uncertainties of the entitlement process in an interesting way. In some places, developers must negotiate a distinct set of payments for each project. Certain constituencies in these communities will oppose a project unless they receive sufficient payments or concessions. In some cities, these payments tend to go to affordable housing; in others they might take the form of labor union contracts or local hire preferences or even private legal settlements. Activists and politicians have developed effective methods for extracting these so-called "community benefits" from housing developments on a project-by-project, ad-hoc basis; and for these activists and politicians, it is essential to keep the transaction costs and regulatory barriers to housing high in order to increase their bargaining power with developers. To the activists fighting for these concessions, it is self-evident that they should try to extract as much funding for their priorities as possible, and they rightly point to negotiated deals that yielded public investments that helped people. But of course, at the level of the housing system as a whole, the resulting profound uncertainty about what level of payments will be required becomes one more factor driving up the cost of housing, scaring away potential investors, and reducing overall housing supply.

Who bears the burden of the costs of the fees and exactions on housing development? At the scale of an individual building, developers cannot simply "pass the costs on" to consumers; rational developers will already be charging the maximum the market will bear. Most of the costs of producing housing (materials, labor, capital) are given from the perspective of the developer; fees and exactions are no different. But if the costs of production go up, developers can try to bid less for land. If all the costs of fees and exaction are known in advance of a land transaction, developers should not bid more than they can afford—which in theory would drive down residual land value.

But there are significant limits, especially in high-demand markets. For one thing, if the rules are inherently unpredictable and changeable, it is nearly impossible to bid rationally on land, which inevitably drives up the cost of capital, and results in inefficient outcomes. More importantly, as a residential developer's offer price decreases, fewer land-sellers will sell, which translates into a reduction in how many parcels will be developed. After all, urban land has other uses than housing. Almost always, the urban parcel in question is generating revenue already; it is occupied by a store, a parking lot, or some other business. It's quite easy to impose such high costs that developers will not be able to outbid existing uses and redevelop so-called

"soft" sites. At some point, the capitalized net operating income flowing from a single-story strip mall retail development is worth more than a housing developer can offer. In jurisdictions like California, this problem is particularly acute because the ballot Proposition 13 approved in 1978 depresses property taxes on long-term owners, further disincentivizing the sale of their existing revenue-generating assets. In the long run, we can expect fees, exactions, and other financial requirements to reduce the quantity of land that is developed. Said differently, the market price for housing has to remain high enough to cover the cost of the fees and exactions, so these function as a price floor that keeps housing more expensive than it otherwise would be.

Most public officials would state that affordable housing is one of their top priorities. But when looking at the combination of housing policies—both the official "affordable housing" policies and the broader set of exclusionary land use regulations—it seems clear that de facto housing policy for most of the cities in expensive metro areas is to make people live somewhere else (and suffer long commutes) or to discourage people from moving into the area in the first place (effectively preventing them from participating in the most successful economies of the country). Many more people experience this sort of exclusion than actually receive a price-restricted, subsidized housing unit, a rent controlled unit, or a housing voucher. The exclusionary effects of unnecessarily high housing costs due to local barriers to supply far outweigh the gains in housing access provided from the other programs. A policy trade-off arises here: is it worth helping one set of lower-income households by providing subsidized housing at the cost of increasing the price of units in the market sector?

For the country as a whole, the restrictive housing policies of the cities in expensive metro areas leads to the segregation of the wealthy into zoned enclave communities; a reduced ability of lower-income people to move to areas of higher opportunity; a diversion of enormous wealth into rent-seeking behavior by landowners; and a decrease in economic productivity for the country as a whole, because labor is not able to be allocated to the most productive economic clusters (Furman 2015; Hseih and Moretti 2015; Gyourko and Molloy 2014; Ganong and Schoag 2015).

The Collective Action Problem of Local Housing Policy

We can understand the tendency for misregulation of the housing market as the result of two sets of factors: first, the jurisdictional fragmentation of American metropolitan areas coupled with the local need to raise money for public services; and second, the combination of locating responsibility for development regulation with localities coupled with control of local democratic process by incumbent homeowners.

Jurisdictional Fragmentation and Local Taxation

Conceptually, both labor markets and housing markets exist at the metropolitan scale, which can be thought of as the "commute shed." Each metropolitan area is

comprised of many individual cities, towns, villages, townships, and usually multiple counties—in other words, local governments that have control over land use decisions. In addition, some regions consist of adjacent and partially overlapping labor markets, which adds further complications—for example, the many cities along the Boston-to-Washington corridor, or the twin and increasingly merged economies of San Francisco and Silicon Valley (Savitch and Adhikari 2017).

Cities compete with one another to avoid "bads" like freeways, dumps, or other land uses with negative local impacts, and also to provide amenities that will be attractive to residents. Competition between cities is supposed to allow citizens to "vote with their feet" to live where they can find the mix of taxes and services that best matches their preferences (Tiebout 1956). While acknowledging that this sorting results partially from divergent personal preferences, it's clear that the outcomes are not all benign. They include the secession of the wealthy into enclaves where they can provide good schools for their children; the segregation of the poor into cities that lack the resources to pay for adequate public services; and a chronic tendency to underproduce housing.

Each city has a fiscal incentive to minimize costs and maximize revenues. Typically that means trying to attract jobs while not adding residents (it is residents who consume public services). Also, each city has an incentive to avoid the negative impacts, especially traffic, that typically come from added housing. Because there are typically many cities within a metropolitan area, it is very possible for some cities to win this fiscal arms race by having a higher ratio of jobs to housing units, enabling those cities to provide higher levels of public service at a lower cost to residents.

From a macro policy perspective, it's not essential for every city in a metropolitan area to produce housing so long as the total housing supply in aggregate is sufficient. But we face pervasive free-rider incentives, which lead every city (technically the people who run the city) to believe it could not possibly be asked to add housing, especially not at high densities, while believing that other cities would be much more logical places to put new housing. Jurisdictional fragmentation at the regional scale coupled with local taxation as the source of funding for essential public services sets up a classic collective action problem.

Localized Control over Land Use and Homeowner Control over Cities

Many things that bear on housing markets are beyond the control of cities: the occupational structure of the economy and the mix of employment opportunities for residents; the distribution of wealth, with all that it implies for purchasing power in the housing market; the expenditure priorities of federal housing and social welfare programs; and so much else. But one thing cities do control in the American system is land use. While there are certain limitations and exceptions (more on these below), the states have delegated land use regulatory power to cities, which exercise that authority through zoning and other development controls. The courts also tend to defer to the judgment of locally elected legislative bodies.

At the same time, smaller cities, comprising most of the land within a metro area, are generally controlled by homeowners because most voters are homeowners

(Jurjevich and Keisling 2015). It does not take a great leap to realize that most voters in most cities are going to be interested in protecting the value of their primary asset (Hertz 2016). In the strong version of this "home voter" hypothesis (as named by Fischel 2001), voters work to suppress housing supply as a way to protect higher housing values. This pattern appears to be especially pervasive in the suburbs. But we can construct a weaker version of the hypothesis, which simply asserts that home-owning voters are not strongly motivated to add supply because housing unaffordability does not directly hurt them, so other factors like the desire to avoid traffic or the desire to protect the character of their neighborhoods outweigh the appeal of seeking to reduce housing costs for other people. In both cases, we would expect that the electoral process would, on average, lead to the selection of politicians who reflect the preferences of their constituents not to add housing.

What about the people who are not homeowners—why are the concerns of renters not showing up in the form of more pro-housing politics? One reason is that most of them do not live in the jurisdiction. Most of the people who would potentially benefit from solving the housing shortage are the ones who have been kept out of the expensive cities to begin with: the people who would be residents, who would not live so far away, or who would join the successful economic cluster, if they were able to. Our local democratic process does not take their interests into account because only people who have already made it "in" are members of the polity.

But even renters in the expensive cities—the people who may or may not occupy a rent-controlled unit, the people who are most at risk of being displaced by rising housing costs—are not always a political force in favor of more open housing markets (Hankinson 2017). This fact is essential for understanding housing politics in the majority-renter cities like New York and San Francisco, and is probably the most difficult aspect of local politics for economists to understand. We have to start by remembering that in many situations, not just housing, people may not be rational about their own self-interest, and may be motivated by things other than self-interest. But we can add nuance to this observation in several ways that make it more understandable why renters might be skeptical about housing development.

In the cities with rent control, plenty of renters have incomes that are so low that they would not be able to afford market prices in any plausible scenario of supply increase. Some have occupied their units for a long time, with rents pegged to much lower levels from years ago. These tenants may be correct in their belief that nothing that adds to the market-rate housing supply will directly help them.

Some tenants fear that new housing development in a previously affordable neighborhood could actually *raise* the prices on the adjacent housing stock—when "gentrification" increases the amenity value of the block, or even by signaling that a street is now "safe" for middle-income residents. My own judgment is that these localized effects are tiny when compared to the overall pricing pressure from regional undersupply, but this is a real debate in many of these cities.

While I think it's clear that the opposition to market-rate housing supply by certain political constituencies inside cities has the effect of enriching homeowners and making the broader housing supply expensive, it is also true that there are localized impacts on particular people that we need to take seriously if we want to change this dynamic (Jacobus 2016). Those who would wish to actually bring down housing costs for everyone and make successful American cities more open once again can't just shrug in response to the displacement of particular individuals and say that nothing can be done; we need a response to the displacement of particular individuals beyond simply shrugging that nothing can be done if we want those individuals (and the leaders who speak on their behalf) to rethink their housing politics. There is a critical role for protecting current residents from displacement by rising housing costs, even while we work to fix the overall housing market.

Political Coalition-Building

Finally, to understand local policy making, we need to pay attention to the strategies pursued by activists and elected officials, who are working to assemble political coalitions. To wield political power it is always necessary to bring together multiple groups of people who have distinct interests and understandings: Judd and Swanstrom (2015) tell the story of changing political coalitions in American cities. Until the 1970s, "growth machine" coalitions of labor unions and business leaders wielded significant clout in many cities, and they still do in some. But antigrowth political coalitions are now widespread.

Renters who fear *increases* in housing prices can be brought into coalition with homeowners who fear *decreases* in housing prices around a shared distrust of elites and a fear of change. But at least in theory, renters who favor lower rents could also be brought into a different coalition with labor unions who favor building, environmentalists who prefer greater density to reduce emissions of greenhouse gases, and immigration rights advocates who believe that making a city more affordable will open it up to new entrants. Both types of political coalitions, and many others, are possible from the same set of interests (Been, Madar, and McDonnell 2014). Perhaps we need more comparative political science research on the formation of divergent urban coalitions, in order to understand why cities have evolved the way they have. But it's clear that the strategies of the political actors matter.

For all of these reasons, we have arrived at a situation in which, to varying degrees, cities in the most economically successful metro areas have systematically created a scarcity of housing. We can understand the undersupply of housing as a logical outcome of the structure of our political system, which combines jurisdictional fragmentation, competition between cities, local control over land use, and control of the city politics by incumbent homeowners. But we also have to give some causal credit in many of these cities to the leaders of what we can call the neighborhood preservation movement, who have managed to build powerful political coalitions that lock in their privilege (Schneider and Teske 1993).

Toward a Better Housing Policy

Many useful changes to housing policy could be made at the national level, encompassing funding for social housing and vouchers, limits on the exclusionary behavior of cities, and more effective forms of social insurance (for some ideas along these lines, see Glaeser and Gyourko 2008). But failures at the national level do not excuse other failures at the local level. Cities are making things worse than they have to be and failing to solve the problems that they could solve. The good news is this: solutions are available that could substantially address the problem of high housing costs. Here are seven ideas.

1. Upzone

The most basic thing that expensive cities need to do to bring down housing costs is to change their zoning to allow more housing to be built, either allowing taller buildings or greater densities or both—in other words, upzoning. Generally, the right way to do this is through careful neighborhood planning to ensure good design and to ensure that we are building complete neighborhoods. The planning process will typically include public realm improvements and infrastructure improvements, not just private buildings. Occasionally there will be major sites that become available such as old shopping malls or industrial sites. More often, new development will be on smaller parcels. The upzoning will be most effective if it is done by many cities across a metropolitan area; and if the process of getting permission to build within the zoning is straightforward and transparent.

Reforming housing policy does not mean getting rid of all regulations.[5] We care about city building for many noneconomic reasons that show up in land use regulations: we want our communities to be beautiful, to nurture a sense of belonging, to express the aspirations of our civilization. We will continue to try to address the sins of our country's past and present racial inequality through land use policies that we hope can help ameliorate segregation.

But there are also many bad reasons to regulate housing. These include the desire to exclude outsiders, the desire to exclude people of a lower socioeconomic status, and a pervasive and understandable desire by incumbent homeowners to protect the value of their properties by preventing changes that they consider undesirable. The solution is not to naively wish for an unregulated housing market; we must instead try to implement a better set of regulations.

[5]Building codes are justified because of the information asymmetry between sellers and buyers, assuring housing purchasers of the safety of the dwelling units they want to occupy. And planning regulations are justified for many reasons that economists should find compelling, including: externalities of property values (what happens on one property can raise or lower the values of adjacent parcels); externalities of environmental costs (settlement patterns determine how much air pollution and greenhouse gasses are generated from transportation); and externalities of public infrastructure (typically, private development is facilitated by public investments in transportation access, water supply, and other infrastructure systems).

2. Rethink Minimal Standards

To reduce the production costs of housing, cities are going to need to look for ways to eliminate some of the regulations that are less essential. That doesn't mean compromising health and safety. It means legalizing smaller units created from accessory dwelling units (a small dwelling that is part of or attached to an existing structure) or single-room occupancy apartments, as well as steps like eliminating parking requirements and looking for ways to encourage innovation in construction techniques (such as prefabricated housing).

There is reason to be skeptical about the ability of public policy changes to reduce production costs of constructing housing. For almost a century, planners have dreamed of applying the techniques of mass production and automation to housing to lower the per-unit construction costs. So far, these dreams have not yielded meaningful results. True mass production should be more possible in greenfield locations, but even here we find a building industry that has not driven costs lower over the decades. It must be difficult to do so.[6] But we should do everything possible to support innovation to reduce the cost of production, and certainly work to remove barriers to lower-cost production techniques, wherever we can (Galante, Draper-Zivetz, and Stein 2017).

3. Connect Superstar Cities to Less-Expensive Places

If people have good transportation access, they can live someplace relatively more affordable and still participate in the economy and social life of a nearby city. In some situations, we can connect communities with less-expensive housing to the cities with the best job markets. Let's call this "the New York model" in honor of the web of rail lines that connects the economic center of Manhattan with towns and cities in every direction, from Philadelphia to Newark to Long Island. This strategy tends to be more available for East Coast cities, which have an inheritance of both rail lines and pre-war, compact towns. It is promising to see that western cities like Denver, Los Angeles, and Seattle have essentially built whole new regional rail networks from scratch over the past decade. Yet even when transit can be created, the transit-supportive, relatively affordable communities do not exist in as large a supply in western cities. In some cases, especially in the West, new transit will make a much bigger difference for housing costs only if it is accompanied by new development.

4. Build More Cities

This is probably the most controversial recommendation on the list from the perspective of city planners. For a century, city planners have debated the idea of "new towns" as a strategy for managing population growth (Fulton 2002; Hall and

[6]One reason is that the housing industry has a lot of inertia. The boom–bust cycle of the real estate economy leads to chronic labor shortages as workers must exit the industry during recessions, while the high cost of housing itself becomes a driver of high wages necessary to attract construction workers—a self-reinforcing cycle in which high housing costs keep housing costs high. And finally, the process of inserting new buildings into the existing urban fabric is by its nature an intricate endeavor.

Ward 1998). Many new towns have been built around the world and in the United States. Unfortunately, most have resulted in highly inefficient land use patterns, high rates of car dependency, and lack of real access to the job-rich city. But the story is not over yet. If sites can be found that are truly within reasonable commuting distance of the jobs in a high-demand city, and if the land can truly be developed at densities equal to traditional cities, it is probably worth experimenting with new cities, to see if we can rediscover the lost art of building great urban places (Duany and Plater-Zyberck 2006; Duany, Speck, and Lydon 2010).

5. Pool Taxes Regionally

We have seen that one of the drivers of housing undersupply is fiscal competition between cities for sales tax and business tax revenues. One structural solution to this problem is to pool sales tax revenues regionally and then redistribute them on a per-capita basis. This is exactly what the Minneapolis metropolitan area does. Its tax-sharing system deserves to be more broadly replicated around the country (Orfield 2002; Orfield and Luce 2010).

6. Move Responsibility for Housing to a Higher Level of Government

We are going to have a much harder time addressing the problem of high housing costs if we continue to defer all land use decisions to the local level. There are simply too many incentives for each jurisdiction to shirk its housing responsibilities and hope that other cities in the region pick up the slack. Portland, Oregon, has a directly elected regional government (called "Metro") that allocates growth to cities within the region as a way to comply with the state's strong growth management law (Abbott 2000). The State of Washington has largely copied Oregon's growth management law, to good effect. Massachusetts has set up a legal process to override local zoning and approve housing developments in jurisdictions that do not comply with state affordable housing requirements (Reid, Galante, and Weinstein-Carnes 2016). In all of these models, the state government has acted to ensure adequate housing supply, recognizing that the incentives and spillover effects of local land use are producing pernicious results. Other states could enact similar reforms.

7. Spend More on Social Housing

Greater spending on social housing should be viewed as a long-term strategy that will help some of the most vulnerable people who are being priced out of expensive cities today. Over time, there are significant upsides to having some portion of urban land be owned by nonmarket actors; it is one tool for reducing the rent-seeking behavior that is channeling so much of the wealth of the most productive cities into a land-owning rentier class. Social housing, just like market-rate housing, is a way to add to the overall supply. Cities and states should experiment with vouchers and new types of delivery mechanisms. Social housing programs do not need to be confined to the same low-income households that today's programs serve; new programs would provide social housing to a broader cross-section of the population as in the European models. In general, funding for these programs

should come from the broad tax base rather than exactions on new housing development in order to avoid the unintended consequences of reducing aggregate housing supply. Recognizing that most people will still obtain their housing in the market, and that we cannot solve the overall problem without a primary emphasis on overcoming the broader housing shortage, there is still an essential role for public spending on social housing.

It may be possible at the scale of the city or the metro area to construct "grand bargains" that include many of these ideas simultaneously. (Seattle's Housing and Livability Agenda, agreed on in 2016, is a possible example.) The good news is that progress on housing prices in the expensive metros is possible, if the political will exists.

Conclusion

A group of metro areas in the United States is simultaneously enjoying both considerable economic success and unprecedented challenges with housing costs. Opening up these metro areas so that far more people can participate in their economic success will require substantial changes to the institutional and physical structure of these metro areas. I have argued that while we can and should spend more money on subsidies for social housing in various forms, this solution cannot scale to help most people. Instead, we will need to do the hard work of reforming our housing markets so that the supply of housing can expand more easily. In other words, we need to change the spatial settlement patterns of the metropolitan areas by adding density within the existing urbanized fabric and/or by creating new urban fabric that is linked by high-quality transportation.

How do we know when we have created "enough" capacity for housing? The per-unit price of land offers one key indicator. What developers call the "pad cost"—the land component of each new housing unit—is the measure of how much restrictive zoning has allowed land owners to capture rents. In the expensive-housing cities, pad costs typically range between $80,000 to $100,000, whereas in the unconstrained sunbelt cities, pad costs are more like $20,000 to $30,000. In the most restrictive zoning regimes, they can rise above $150,000. When a city has zoned for sufficient capacity, bidders on land have many options for which parcels to purchase. Of course, the price per square foot of land cannot go lower than the other available uses of the land: a developer generally has to buy out the business that operates on the site—the store, the parking lot, or whatever it may be. But if a site is zoned for very high densities, and if many sites all over the city are zoned for very high densities, then the per-unit cost of land can be driven quite low. Indeed, it would be useful to have public agencies, or maybe even researchers at the regional Federal Reserve banks, track the per-unit cost of land and other indicators as a guide to housing policy.

Might we reach a point where a city is "full" and cannot (or should not) add population? This has been an important debate in planning theory (Lynch 1981). Physically, the answer is "no." We observe a great range of settlement and urban

density patterns across the world, and US cities are not especially dense. Moreover there are great ecological benefits to increasing the density of US settlement patterns as a way to reduce per capita energy consumption (Newman and Kenworthy 1999). The most relevant limits to growth in a metropolitan area are political and aesthetic, not physical.

We will lift far more people into the middle class if we can make it easier to join successful urban economies than it is today. In addition, we will reduce the ecological footprint of our nation if we make it easier for urban growth to happen in compact forms rather than in sprawling suburban patterns. The solution to high housing costs in the expensive metro areas of the United States is also a solution for increasing economic opportunity and increasing ecological resilience.

■ *I would like to thank the following people for reviewing drafts of this essay and providing helpful commentary: Joe Cortright, Kim-Mai Cutler, Ted Egan, Rick Jacobus, Chris Jones, Sarah Karlinsky, Steve Levy, Sharon Metcalf, Doug Shoemaker, Randy Smith, Michael Teitz, Steve Waldman, Michael Yarne, and Sarah Jo Szambelan. All responsibility for the content is of course my own.*

References

Abbott, Carl. J. 2000. "The Capital of Good Planning: Metropolitan Portland since 1970." In *The American Planning Tradition: Culture and Policy*, edited by Robert Fishman, 241–61. Washington, DC: Woodrow Wilson Center Press.

Apgar, William C., Jr. 1990. "Which Housing Policy is Best." *Housing Policy Debate* 1(1): 1–32.

Arnott, Richard. 1995. "Time for Revisionism on Rent Control?" *Journal of Economic Perspectives* 9(1): 99–120.

Arnott, Richard J., and Joseph E. Stiglitz. 1979. "Aggregate Land Rents, Expenditure on Public Goods, and Optimal City Size." *Quarterly Journal of Economics* 93(4): 471–500.

Bauer, Catherine. 1934. *Modern Housing.* Boston and New York: Houghton Mifflin Company.

Been, Vicki, Josiah Madar, and Simon McDonnell. 2014. "Urban Land-Use Regulations: Are Homevoters Overtaking the Growth Machine?" *Journal of Empirical Legal Studies* 11(2): 227–65.

Center on Budget and Policy Priorities. 2017. "Policy Basics: Federal Rental Assistance." May 3. https://www.cbpp.org/research/housing/policy-basics-federal-rental-assistance.

Collinson, Robert, and Peter Ganong. 2017. "How Do Changes in Housing Voucher Design Affect Rent and Neighborhood Quality?" June. http://furmancenter.org/files/collinsonGanong0502.pdf.

DeLuca, Stefanie, and Peter Rosenblatt. 2017. "Walking Away from *The Wire*: Housing Mobility and Neighborhood Opportunity in Baltimore." *Housing Policy Debate* 27(4): 519–46.

Duany, Andres, and Elizabeth Plater-Zyberck. 2006. *Towns and Town-Making Principles.* New York: Rizolli.

Duany, Andres, Jeff Speck, and Mike Lydon. 2010. *The Smart Growth Manual.* New York: McGraw-Hill.

Economist, The. 2015. "The Paradox of Soil." April 4. http://www.economist.com/news/briefing/21647622-land-centre-pre-industrial-economy-has-returned-constraint-growth.

Ellen, Ingrid Gould, and Brendan O'Flaherty. 2013. "How New York and Los Angeles Housing Policies are Different—And Maybe Why" In *New York and Los Angeles: The Uncertain Future*, edited by David Halle and Andrew A. Beveridge, 286–309.

New York: Oxford University Press.

Erickson, David J. 2009. *The Housing Policy Revolution: Networks and Neighborhoods*. Washington, DC: Urban Institute Press.

Fischel, William A. 2001. *The Homevoter Hypothesis; How Home Values Influence Local Government Taxation, School Finance, and Land-Use Policies*. Cambridge, MA: Harvard University Press.

Fischel, William A. 2015. *Zoning Rules: The Economics of Land Use Regulation*. Cambridge, MA: Lincoln Institute of Land Policy.

Fischer, Will, and Barbara Sard. 2016. "Chart Book: Federal Housing Spending is Poorly Matched to Need." Center on Budget and Policy Priorities, June 8, http://www.cbpp.org/research/housing/chart-book-federal-housing-spending-is-poorly-matched-to-need.

Fulton, William. 2002. "The Garden Suburb and the New Urbanism." In *From Garden City to Green City: The Legacy of Ebenezer Howard*, pp. 159–70. Baltimore: John Hopkins University Press.

Furman, Jason. 2015. "Barriers to Shared Growth: The Case of Land Use Regulation and Economic Rents." Remarks ("expanded version of the remarks as prepared for delivery"). November 20. https://obamawhitehouse.archives.gov/sites/default/files/page/files/20151120_barriers_shared_growth_land_use_regulation_and_economic_rents.pdf.

Galante, Carol, Sara Draper-Zivetz, and Allie Stein. 2017. "Building Affordability by Building Affordably: Exploring the Benefits, Barriers, and Breakthroughs Needed to Scale Off-Site Multifamily Construction." Terner Center for Housing Innovation, UC Berkeley. March. http://ternercenter.berkeley.edu/uploads/offsite_construction.pdf.

Ganong, Peter, and Daniel Shoag. 2015. "Why has Regional Income Convergence in the U.S. Declined?" January. http://scholar.harvard.edu/files/shoag/files/why_has_regional_income_convergence_in_the_us_declined_01.pdf.

Glaeser, Edward L., and Joseph Gyourko. 2008. *Rethinking Federal Housing Policy: How to Make Housing Plentiful and Affordable*. Washington, DC: American Enterprise Institute.

Groth, Paul. 1994. *Living Downtown: The History of Residential Hotels in the United States*. Berkeley: University of California Press.

Gyourko, Joseph, and Raven Molloy. 2014. "Regulation and Housing Supply." NBER Working Paper 20536, October.

Gyourko, Joseph, Christopher Mayer, and Todd Sinai. 2013. "Superstar Cities." *American Economic Journal: Economic Policy* 5(4): 167–99.

Haas, Peter M., Carrie Makarewicz, Albert Benedict, Thomas W. Sanchez, and Casey J. Dawkins. 2006. "Housing & Transportation Cost Trade-offs and Burdens of Working Households in 28 Metros." Centre for Neighbourhood Technology, July.

Hall, Peter, and Colin Ward. 1998. *Sociable Cities: The Legacy of Ebenezer Howard*. John Wiley and Sons.

Hankinson, Michael. 2017. "When Do Renters Behave Like Homeowners? High Rent, Price Anxiety, and NIMBYism." Working Paper, Harvard Joint Center for Housing Studies. February. http://www.jchs.harvard.edu/sites/jchs.harvard.edu/files/harvard_jchs_hankinson_2017_renters_behave_like_homeowners_0.pdf.

Hertz, Daniel. 2016. "Housing Cannot Be a Good Investment and Affordable." *City Observatory*, July 20. http://cityobservatory.org/housing-cant-be-a-good-investment-and-affordable.

Hickey, Robert, Lisa Sturtevant, and Emily Thaden. 2014. "Achieving Lasting Affordability through Inclusionary Housing." Lincoln Institute of Land Policy Working Paper. http://www.inhousing.org/wp-content/uploads/PollockPRisingInterestInclusionaryHousingLincolnInstituteWorkingPaper.pdf.

Housing Europe. 2015. *The State of Housing in the EU 2015*. http://www.housingeurope.eu/resource-468/the-state-of-housing-in-the-eu-2015.

Hsieh, Chang-Tai, and Enrico Moretti. 2015. "Housing Constraints and Spatial Misallocation." NBER Working Paper 21154.

Jacobus, Rick. 2015. "Inclusionary Housing; Creating and Maintaining Equitable Communities." Lincoln Institute of Land Policy. http://www.inhousing.org/wp-content/uploads/Inclusionary-Housing-Report-2015.pdf.

Jacobus, Rick. 2016. "Why We Must Build." *Shelterforce,* March 10. https://shelterforce.org/2016/03/10/why_we_must_build/.

Joseph, Mark L., Robert J. Chaskin, and Henry S. Webber. 2007. "The Theoretical Basis for Addressing Poverty Through Mixed-Income Development." *Urban Affairs Review* 42(3): 369–409.

Judd, Dennis R, and Todd Swanstrom. 2015. *City Politics,* Ninth Edition. New York: Routledge.

Jurjevich, Jason R., and Phil Keisling. 2015. "Who Votes for Mayor?" PSU Pilot Research Report, Portland State University, July. Knight Foundation.

Lynch, Kevin. 1981. *Good City Form*. Boston: MIT Press.

Metcalf, Gabriel. 2015. *Democratic by Design: How Carsharing, Co-ops, and Community Land Trusts are Reinventing America*. New York: St. Martin's Press.

Newman, Peter, and Jeffrey Kenworthy. 1999. *Sustainability and Cities*. Washington, DC: Island Press.

Orfield, Myron. 2002. *American Metropolitics: The New Suburban Reality*. Washington, DC: Brookings Institution Press.

Orfield, Myron, and Thomas F. Luce, Jr., eds. 2010. *Region: Planning the Future of the Twin Cities*. University of Minnesota Press.

O'Sullivan, Feargus. 2016. "The Rent is Now Somewhat Less High in Paris." In *The Atlantic CityLab*, August 3.

Radford, Gail. 1996. *Modern Housing for America: Policy Struggles in the New Deal Era*. University of Chicago Press.

Reid, Carolina, Carol Galante, and Ashley Weinstein-Carnes. 2016. "Borrowing Innovation, Achieving Affordability: What We Can Learn from Massachusetts Chapter 40B." Policy Paper 1, August, Terner Center for Housing Innovation, University of California, Berkeley. http://ternercenter.berkeley.edu/uploads/California_40B_Working_Paper.pdf.

Rognlie, Matthew. 2015. "Deciphering the Fall and Rise in the Net Capital Share." *Brookings Papers on Economic Activity,* Spring.

San Francisco Office of the Controller. 2016. "Inclusionary Housing Working Group: Final Report." February 13, http://sfcontroller.org/sites/default/files/Documents/Economic%20Analysis/Final%20Inclusionary%20Housing%20Report%20February%202017.pdf.

Savitch, H. V., and Sarin Adhikari. 2017. "Fragmented Regionalism: Why Metropolitan America Continues to Splinter." *Urban Affairs Review* 53(2) 381–402.

Schleicher, David. 2013. "City Unplanning." *Yale Law Journal* 122(7): 1670–1737.

Schneider, Mark, and Paul Teske. 1993. "The Antigrowth Entrepreneur: Challenging the 'Equilibrium' of the Growth Machine." *Journal of Politics* 55(3): 720–36.

Schwartz, Alex F. 2015. *Housing Policy in the United States,* Third Edition. New York: Routledge.

Smith, Doug. 2017. "Up to 600,000 Expected to Apply When L.A. Reopens Section 8 Housing List This Month after 13 Years." *Los Angeles Times,* October 1.

Solomon, Daniel. 2003. *Global City Blues*. Washington, DC: Island Press.

Stone, Michael E., Terry Burke, and Liss Ralston. 2011. "The Residual Income Approach to Housing Affordability: The Theory and the Practice." http://works.bepress.com/michael_stone/7.

Sturtevant, Lisa A. 2016. "Separating Fact from Fiction to Design Effective Inclusionary Housing Programs." Inclusionary Housing, Research and Policy Brief, Center for Housing Policy, http://media.wix.com/ugd/19cfbe_9a68f933ed6c45bfb5f8b7d2ef49dda0.pdf.

Tiebout, Charles M. 1956. "A Pure Theory of Local Expenditures." *Journal of Political Economy* 64(5): 416–24.

Williams, Stockton, Ian Carlton, Lorelei Juntunen, Emily Picha, and Mike Wilkerson. 2016. "The Economics of Inclusionary Development." Washington, DC: Urban Land Institute. http://uli.org/wp-content/uploads/ULI-Documents/Economics-of-Inclusionary-Zoning.pdf.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 81–96

Friedman's Presidential Address in the Evolution of Macroeconomic Thought

N. Gregory Mankiw and Ricardo Reis

Presidential addresses to the American Economic Association are always notable events. They are given by scholars of great repute who, by virtue of their office, are being honored by the broad economics profession. The talks are attended by large crowds at the annual AEA meetings. They are prominently published as the lead article in an issue of the *American Economic Review*, one of the discipline's most widely read journals. It is no surprise, therefore, that these addresses often play a significant part in the evolution of the field.

Milton Friedman's presidential address, "The Role of Monetary Policy," which was delivered 50 years ago in December 1967 and published in the March 1968 issue of the *American Economic Review*, is nonetheless unusual in the outsized role it has played. Citation counts offer one measure of its influence. As of this writing, the article has been cited more than 7,500 times according to Google Scholar, making it the third most-cited presidential address in AEA history, beaten only by the addresses of Simon Kuznets on "Economic Growth and Income Inequality" (delivered in 1954, published in 1955) and Theodore Schultz on "Investment in Human Capital" (delivered in 1960, published in 1961). Friedman's address is cited less than his 1962 book *Capitalism and Freedom* and less than a brief essay he wrote in *The New York Times Magazine* in 1970, "The Social Responsibility of Business Is to

■ *N. Gregory Mankiw is Robert M. Beren Professor of Economics at Harvard University, Cambridge, Massachusetts. Ricardo Reis is A. W. Phillips Professor of Economics, London School of Economics and Political Science, London, United Kingdom. Their email addresses are ngmankiw@harvard.edu and r.a.reis@lse.ac.uk.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.1.81

doi=10.1257/jep.32.1.81

Increase Its Profits." But the citation count for Friedman's presidential address is roughly on par with the 1963 *A Monetary History of the United States* by Friedman and Anna Schwartz. Aside from these counterexamples, the 1967 presidential address is cited more often than anything else Friedman wrote during his long, prolific, and influential career.

What explains the huge influence of this work, merely 17 pages in length? One factor is that Friedman addresses an important topic. Another is that it is written in simple, clear prose, making it an ideal addition to the reading lists of many courses. But these same points can be made for many other AEA presidential addresses. What distinguishes Friedman's address is that it invites readers to reorient their thinking in a fundamental way. It was an invitation that, after hearing the arguments, many readers chose to accept. Indeed, it is no exaggeration to view Friedman's 1967 AEA presidential address as marking a turning point in the history of macroeconomic research.

Our goal here is to assess this contribution, with the benefit of a half-century of hindsight. We discuss where macroeconomics was before the address, what insights Friedman offered, where researchers and central bankers stand today on these issues, and (most speculatively) where we may be heading in the future. We focus on the presidential address alone, putting aside Friedman's many other contributions (discussed, for example, in Nelson 2017).

Macroeconomics before the Address

Let's start by setting the stage. When Friedman gave his address in 1967, one author of the present essay was in grade school and the other was not yet born, so neither of us can claim first-hand experience. But using the historical record, only a little imagination is needed to get a sense of what was occupying the thoughts of most macroeconomists as Friedman walked to the podium.

There seems little doubt that the focal event for macroeconomists of that era was still the Great Depression of the 1930s. By the late 1960s, the Depression, rather than being a recent event, had started to fade into history. (To put it in perspective, the Depression was then about as current as the presidency of Ronald Reagan is today.) But many of the macroeconomists listening to Friedman, especially the more senior ones, had lived through this historic downturn, and it was often the motivating event of their professional lives.

That was surely true for Friedman. In his contribution to the wonderful collection *Lives of Laureates* (edited by Breit and Hirsch 2004), Friedman wrote:

> "I graduated from college in 1932, when the United States was at the bottom of the deepest depression in its history before or since. The dominant problem of the time was economics. How to get out of the depression? How to reduce unemployment? What explained the paradox of great need on the one hand and unused resources on the other? Under the circumstances, becoming an

> economist seemed more relevant to the burning issues of the day than becoming an applied mathematician or an actuary" (pp. 69–70).

Today, we can say with confidence that the world is a better place for Milton Friedman having forgone the opportunity to become an actuary!

In the decades after Friedman graduated from college, economists slowly developed an understanding of how to view fluctuations. That understanding was founded on John Maynard Keynes's landmark book *The General Theory of Employment, Interest and Money* (1936). Keynes's vision was clarified and simplified—some would say oversimplified—in the work of Hicks (1937) and Hansen (1953). Their IS–LM model provided the benchmark theory for explaining how insufficient aggregate demand led to economic downturns, as well as how monetary and fiscal policy could combat those downturns. It also provided the starting point for larger econometric models used for forecasting and policy analysis, such as the Federal Reserve's MPS model, work on which began in 1966 under the leadership of Franco Modigliani, Albert Ando, and Frank de Leeuw. The name MPS is derived from MIT, University of Pennsylvania, and Social Science Research Council (Brayton, Levin, Lyon, and Williams 1997).

The IS–LM model takes the price level as given, which is perhaps a reasonable assumption in the shortest of short runs, but the economists of that era were also concerned about the forces that led the price level to change over time. One important reference is the 1960 paper by Paul Samuelson and Robert Solow, "Analytical Aspects of Anti-Inflation Policy." Samuelson and Solow discuss the many forces that influence inflation, emphasizing the difficulty of identifying whether any rise in inflation is driven by an increase in costs or an increase in demand. However, their essay is probably best remembered for its emphasis on the Phillips curve as a useful addition to the macroeconomist's toolbox. Friedman does not cite this paper in his presidential address, but it is nonetheless representative of the worldview which many mainstream macroeconomists had adopted and to which Friedman was responding.

Samuelson and Solow (1960) presented the Phillips curve as "the menu of choice between different degrees of unemployment and price stability" (p. 192). While the idea of such a menu was their main thrust, they recognized the possibility that it might not be stable over time. In particular, they discussed various ways in which a low-pressure economy—one with low inflation and high unemployment—might shift the Phillips curve over time. On the one hand, "it might be that the low-pressure demand would so act upon wage and other expectations as to shift the curve downward in the longer run" (p. 193). On the other hand, a "low-pressure economy might build up within itself over the years larger and larger amounts of structural unemployment," resulting in "an upward shift of our menu of choice" (p. 193). Thus, Samuelson and Solow anticipated what would later be known as the expectation-augmented Phillips curve and "hysteresis effects" (which refer to the possibility of long-lasting increases in unemployment after a recession). But these effects were considered caveats to their main analysis,

rather than central to it. For most readers of their paper, the main take-away was the Phillips curve as a menu of outcomes available to policymakers, both in the short run and in the long run.

The Key Insights

Enter Milton Friedman's AEA presidential address in December 1967, only a few years after he and Anna Schwartz had published their *Monetary History*. Though Friedman had immersed himself in monetary history, he did not use this opportunity to review the historical record. Instead, the address is largely a work of monetary theory, aimed at providing a big picture view of the potential and limits of monetary policy. It is worth noting that Friedman's perspective echoes certain ideas presented, roughly concurrently, by Edmund Phelps (1967, 1968). It is unclear to us whether Friedman was aware of Phelps's work in this area or whether—what is more likely in light of the fact that neither cited the other—these two great scholars were led in the same direction by the intellectual climate of the time.

A first major theme of Friedman's (1968) address is its focus on the behavior of the economy in the long run. Samuelson and Solow (1960) seemed to view the long run as merely the consequence of a series of Keynesian short runs. In contrast, Friedman (1968) viewed the long run as the timeframe under which we should apply the principles of classical economics, especially monetary neutrality. Regardless of what the central bank did, unemployment would over time approach its natural rate, which he defined as "the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availability, the costs of mobility, and so on" (p. 8). This understanding of how the economy worked in the long run provided the basis for, and restrictions on, how we tried to understand the behavior of the economy in the short run.

A second and related major theme of Friedman's (1968) address is its focus on expectations. As noted, Samuelson and Solow (1960) had previously mentioned the role of expectations, and they understood that it might distinguish the short run from the long run. But this concern was not their main focus, and they attached no particular significance to whether actual and expected inflation are the same. By contrast, for Friedman, expectations were the key to explaining how the economy might appear to face a Phillips curve trade-off and how that trade-off would disappear if we tried to exploit it. He wrote that "there is always a temporary trade-off between inflation and unemployment; there is no permanent trade-off. The temporary trade-off comes not from inflation per se, but from unanticipated inflation, which generally means, from a rising rate of inflation" (1968, p. 11). The deviation of reality from expectations was what permitted the economy to depart from its classical benchmark. But because over time people catch on to what is happening,

expectations and reality must eventually come into line, ensuring that these departures are only transitory.

Friedman's focus on the long run and his emphasis on expectations are closely connected. In some macroeconomic models, the long run is the time horizon over which nominal wages and prices can overcome their short-run stickiness, allowing the economy to return to its classical equilibrium. Friedman, instead, viewed the long run as the time horizon over which people become better informed and so their expectations align with reality.

By bringing expectations to the center of the story, Friedman's address helped to usher in the rational expectations revolution that followed. Influential articles in the 1970s by Lucas (1972), Sargent and Wallace (1975), and Barro (1977) were built on the conceptual foundation that Friedman had put in place. Nonetheless, it is worth noting that Friedman (1968) gave no hint that he thought expectations were as rational as these later authors would assume. Indeed, his emphasis on unanticipated inflation, along with his judgment that it took "something like two to five years" (p. 11) for the real effects to dissipate, suggests that he thought expectations were slow to adapt to changes in the policy environment. While it is possible that he had some other propagation mechanism in mind to explain these persistent effects, the address is most naturally read through the lens of old-fashioned adaptive expectations. From a modern perspective, Friedman's assumption that expectations are sluggish rather than rational seems prescient. As we will discuss shortly, recent research on how people form expectations has moved in this direction.

Implications for Monetary Policy

Using these themes of the classical long run and the centrality of expectations, Friedman takes on policy questions with a simple bifurcation: what monetary policy cannot do and what monetary policy can do. It is a division that remains useful today (even though, as we discuss later, modern macroeconomists might include different items on each list).

Friedman begins with what monetary policy cannot do. He emphasizes that, except in the short run, the central bank cannot peg either interest rates or the unemployment rate. The argument regarding the unemployment rate is that the trade-off described by the Phillips curve is transitory and unemployment must eventually return to its natural rate, and so any attempt by the central bank to achieve otherwise will put inflation into an unstable spiral. The argument regarding interest rates is similar: because we can never know with much precision what the natural rate of interest is, any attempt to peg interest rates will also likely lead to inflation getting out of control. From a modern perspective, it is noteworthy that Friedman does not consider the possibility of feedback rules from unemployment and inflation as ways of setting interest rate policy, which today we call "Taylor rules" (Taylor 1993).

When Friedman turns to what monetary policy can do, he says that the "first and most important lesson" is that "monetary policy can prevent money itself from

being a major source of economic disturbance" (p. 12). Here we see the profound influence of his work with Anna Schwartz, especially their *Monetary History of the United States*. From their perspective, history is replete with examples of erroneous central bank actions and their consequences. The severity of the Great Depression is a case in point.

It is significant that, while Friedman is often portrayed as an advocate for passive monetary policy, he is not dogmatic on this point. He notes that "monetary policy can contribute to offsetting major disturbances in the economic system arising from other sources" (p. 14). Fiscal policy, in particular, is mentioned as one of these other disturbances. Yet he cautions that this activist role should not be taken too far, in light of our limited ability to recognize shocks and gauge their magnitude in a timely fashion.

The final section of Friedman's presidential address concerns the conduct of monetary policy. He argues that the primary focus should be on something the central bank can control in the long run—that is, a nominal variable. He considers the nominal exchange rate, the price level, and monetary aggregates. He says that the exchange rate is not sufficiently important, given the small role of trade in the US economy. While the price level is the most important of these variables, he argues that the link between central bank actions and the price level is too long and unpredictable for the price level to serve as a useful policy target. He concludes that steady growth in some monetary aggregate is the best starting point for policy.

This last recommendation may be the part of Friedman's analysis with which macroeconomists today would most strongly disagree (for an exception, see Hetzel 2017). The economy is subject to many types of shocks, such as oil price changes, financial crises, and shifting animal spirits of investors. In many cases, simply keeping a monetary aggregate on a steady path seems an insufficient response to macroeconomic distress. Moreover, in a world with an increasingly complex array of financial instruments, determining an appropriate measure of the quantity of money to target is difficult and perhaps insuperable. As a result, over the past few decades, the ratio of nominal income to many measures of money (what is called "velocity") has been unstable, convincing most economists and policymakers that targeting money would lead to large fluctuations in prices and incomes.

The Current State of Play

The Great Recession that followed the financial crisis of 2007–2008 may become the defining moment for a new generation of macroeconomists, just as the Great Depression was for Milton Friedman's generation. The initial contraction in production and the turmoil in financial markets were as serious as those in 1929. Like classical economics in the 1930s, which had been criticized for not explaining why so many people who wanted a job could not find one, modern economics was criticized for not forecasting the crash. In a visit to the London School of Economics, the Queen of England famously asked (as reported in Pierce

2008): "Why did nobody notice it?" Macroeconomics responded, and researchers have been fervently at work modeling banks and financial markets, using microeconomic data to better calibrate and estimate models, and studying unconventional monetary policies. The current state of play is not the same as it was ten years ago.

It is a testament to the reach of Friedman's (1968) presidential address that its two main themes—the use of the long-run time frame and the centrality of expectations—remain integral to macroeconomics and have not been greatly affected by the crisis. Most classes in macroeconomics for more than two decades have started with the long run, as many graduate and undergraduate textbooks will testify. Students first learn about the Solow (or Ramsey) models for the evolution of real variables and then use the classical dichotomy and the Fisher equation for interest rates to discuss nominal variables. To be sure, there is greater heterogeneity across institutions and teachers about what models are introduced next. But the starting point, just as in Friedman's address, is almost always a long-run classical benchmark. Keynes (1923) famously wrote: "The long run is a misleading guide to current affairs. In the long run we are all dead." But Friedman won the discussion about the relevance of the long run to current decisions, and economists today work through death before trying to make sense of life.

When Friedman wrote his address, most students organized their thoughts about business cycles using the IS–LM model. This model gives at best a secondary role to expectations. While early Keynesians sometimes emphasized the animal spirits of investors, these were taken to reflect irrational exogenous sentiments rather than purposeful forward-looking behavior. This is far from the reality of modern macroeconomics. Almost all macroeconomic analyses now emphasize intertemporal trade-offs, so the beliefs of economic agents about the future have become a crucial part of the story. Expectations remain at the forefront of macroeconomic analysis, just as Friedman advised.

In particular, modern theories of price dynamics give expected inflation a key role, and in doing so, they embed Friedman's hypothesis that unemployment eventually returns to its natural rate, regardless of the policies pursued by the central bank. To be sure, some researchers have questioned this hypothesis and proposed theories of hysteresis, under which monetary policy can have real effects in the long run. But these arguments are the exception rather than the rule. For most macroeconomists, the natural-rate hypothesis remains the touchstone.

At the same time, the current state of play is also quite different from either the adaptive expectations that Friedman seemed to use or the rational expectations that were at the center of research in the 1970s. With rational expectations, there is, as Sargent (2008) noted, a "communism of beliefs": All economic agents believe the same thing, because they perfectly observe all the same variables and use the exact same model to combine them. This model is the one given to them by the omniscient model-builder. Economic theorists initially embraced this assumption because it offered them an elegant, model-consistent way to treat expectations. However, for several decades now, as expectations have become central not only to policy but also to research in economics, the rationality of

expectations, as conventionally defined, is often called into question. It is common today to sit through seminars in macroeconomics and see presenters assume that the economic agents only imperfectly or infrequently observe some variables, or have limited attention, or learn according to a least-squares formula, or apply other heuristics that are behaviorally founded. Few in the audience wince at seeing these alternatives. Much like the long run, rational expectations may still be the starting point in the classroom, but years of research have produced more nuanced models of how people look into the future.

Expectations are now also central in empirical work. With Justin Wolfers, the two of us made the point long ago that progress in studying expectations required that economists look at microdata from surveys (Mankiw, Reis, and Wolfers 2004). There is a rich amount of panel data reporting people's survey answers to what they expect about numerous variables. While researchers had long looked at the average of these expectations, we emphasized that one should also examine disagreement across people and how it evolves over time. Moreover, researchers can see how individual characteristics, like age or income, might affect the accuracy of these expectations and how often they are updated. In the study of inflation dynamics, many active researchers are using these data to study which of the alternatives to rational expectations should supplant it as the benchmark (for example, Coibion and Gorodnichenko 2012; Malmendier and Nagel 2016; Andrade, Crump, Eusepi, and Moench 2016). There is not yet a consensus about which theory of expectations is most useful, but there is no doubt that expectations data are more central than ever in macroeconomics today, just as Friedman suggested they should be.

Friedman's analysis of macroeconomic fluctuations from the perspective of a Phillips curve that is anchored by the long run is also alive and well. In fact, the last decade has provided a new application of Friedman's logic. Friedman predicted that the Phillips curve that had appeared in the data throughout the 1950s and 1960s would break down if policymakers followed Samuelson and Solow's (1960) advice and started exploiting it. The stagflation of the 1970s, when both inflation and unemployment rose, is one of the greatest successes of out-of-sample forecasting by a macroeconomist. Soon after, macroeconomists could be split into camps of "freshwater" and "saltwater" varieties, in Hall's (1976) famous characterization, depending on the extent to which their theories were anchored by the tenets of classical economics. Yet by the start of this century, macroeconomists had again converged on a view of the trade-off facing central banks that merged the short-run insights from New Keynesian economics summarized in Mankiw and Romer (1991) and the long-run properties of the dynamic general equilibrium models of Kydland and Prescott (1982), as Blanchard (2009) described. In honor of the neoclassical synthesis of Samuelson and Solow, Goodfriend and King (1997) labeled this approach the New Neoclassical Synthesis. From this perspective, Friedman's address can be viewed as a starting point for dynamic stochastic general equilibrium models (though Friedman might well have looked askance at some aspects of DSGE methodology).

At the heart of this new synthesis was a Phillips curve built on the work of Taylor and Calvo (discussed in Taylor 2016). Firms were assumed to set prices equal to the average of their expected future marginal costs, but to alter prices in an infrequent and staggered way. From the start, however, researchers saw flaws in this Phillips curve. Ball (1994) provided a pointed critique of its use for policymaking: He showed that the model predicted that times of announced disinflation should be times of economic expansion, which was almost never true in reality. And, because the firms that are adjusting their prices today respond strongly to future expected events, inflation in the model can jump without any of the inertia observed in the data.

Models in the early 2000s attempted to remedy these problems by assuming that firms partially indexed their prices to lagged inflation. This approach introduced inflation inertia by sheer assumption. Smets and Wouters (2007) found that this model could fit the US data for the previous four decades reasonably well. Yet the empirical success of their model could end up sharing the same fate as that of Samuelson and Solow (1960). Just as Milton Friedman had done before, some researchers suggested that given its shaky foundations, this new Phillips curve was bound to break down, as soon as there was a large shock or a change in policy regime. In Ball, Mankiw, and Reis (2005), we pointed to "the sorry state of monetary policy analysis" and echoed Friedman in writing that "it is imperative that expectations be allowed to adjust to the new regime." The most recent decade of data has provided yet another vindication for Milton Friedman's arguments, as the slope and location of the Phillips curve again shifted, invalidating previous estimates (Coibion and Gorodnichenko 2015; Blanchard 2016).

The Role of Monetary Policy Today

Modern macroeconomics is further from Friedman's views regarding what monetary policy cannot, can, and should do. The belief that, in the long run, the central bank can do little about real variables is still canon for most macroeconomists, and few would suggest that monetary policy should have targets for labor force participation, inequality, or the long-term real interest rate. Yet, it is not uncommon today to hear central bankers pontificate in speeches about such issues. Friedman's example that a speech or article about monetary policy should spend almost as much space on what the central bank cannot do, as it does on what it can do, has eroded over time.

While Friedman favored targeting the growth rate of a monetary aggregate, macroeconomists have for the last two decades instead embraced targets for inflation given to independent central banks (Svensson 2010). The major central banks in the developed economies of the world today all share not just a target for inflation but even a specific number—namely, 2 percent—differing only in how strictly and quickly they strive to achieve it. Friedman worried that it would be hard to hit any target for prices, yet the track record so far has been quite successful, with

annual inflation almost never straying from the band between 0 and 4 percent. For the central bank with the strictest target, the European Central Bank, the price level at the end of 2016 was 38 percent higher than it had been at the end of 1998, when the ECB started operations. An exact target of 2 percent per year would have predicted a 42 percent increase. The annualized deviation from target averages a mere 0.2 percent over this 18-year period for the ECB, a success that Friedman was skeptical could be achieved.

Modern macroeconomics also focuses more on the nominal interest rate than on monetary aggregates, both as an instrument for policy and as a guide to the state of the economy. Friedman's presidential address discussed Knut Wicksell's concept of a natural rate of interest but dismissed it as a good guide for policy. Today and for many years now, Friedman has lost this argument to Woodford (2003), who convinced academics and central bankers to embrace the Wicksellian use of interest rates as the main policy tool and their deviation from natural rates as the key policy target. The central bank directly controls one interest rate, and the effect of its actions on other interest rates is measured more reliably than the effect on money. Moreover, there is a clear link from interest rates to the price of credit and to the willingness of people to save or borrow. The Federal Reserve unequivocally states that "the importance of the money supply as a guide for the conduct of monetary policy in the United States has diminished over time" (in the FAQ section of its website at https://www.federalreserve.gov/faqs/money_12845.htm).

Friedman recommended strict rules to guide monetary policy because he thought that deviating from such rules added noise into the system, leading to inefficient fluctuations in inflation and the real economy. Many modern macroeconomists seem to agree, given the paucity of academic or applied arguments in defense of purely discretionary choices by central bankers. Chari and Kehoe (2006), summarizing in this journal the modern study of commitment and the potential time inconsistency of discretionary policy, emphatically wrote: "The message of examples like these is that discretionary policy making has only costs and no benefits, so that if government policymakers can be made to commit to a policy rule, society should make them do so." At the same time, almost no central bank has adopted a strict rule for monetary policy. Instead, central banks have continued to use a great deal of discretion to infer the state of the economy from many imperfect measures, and to react to the wide variety of shocks. However, policymakers have responded to academics by placing a large emphasis on the transparency of central bank actions. Central bank governors give frequent speeches, their institutions publish detailed reports justifying their actions, and academic research has taken this transparency as given, busying itself instead with how to shape and conduct central bank communication (Blinder, Ehrmann, Fratzscher, De Haan, and Jansen 2008). Such efforts at transparency can be seen as trying to reduce the noise arising from central bank actions.

At the same time, modern central banks interpret inflation targets in a flexible way, with a willingness to trade off deviations of inflation from target against movement in real activity (Woodford 2010). By following feedback rules that condition

policy on the state of the business cycle, central banks aggressively respond to recessions and booms and thus explicitly commit to the countercyclical stabilization policies that Friedman thought were fruitless. Galí and Gertler (2007) in this journal characterized the two insights of modern macroeconomic models for monetary policy as being: "1) the significant role of expectations of future policy actions in the monetary transmission mechanism and 2) the importance for the central bank of tracking the flexible price equilibrium values of the natural levels of output and the real interest rate." Friedman would have applauded the first, but the second goes against the main thrust of the policy recommendations in his presidential address.

Moreover, Friedman's presidential address argued that "too late and too much has been the general practice" of monetary policy because of "the failure of monetary authorities to allow for the delay between their actions and the subsequent effects on the economy" (p. 16). Modern central banks agree but have responded by adopting a policy of "inflation forecast targeting" (Woodford 2007): that is, they discuss their policies in terms of what will bring forecasted inflation two or three years ahead back on target.

Finally, the Great Recession and the actions of the Federal Reserve provide a useful contrast between the central bank that Milton Friedman wished for and the one that exists today. Friedman (p. 14) thought that "monetary policy can contribute to offsetting major disturbances in the economic system arising from other sources," but he says that "I have put this point last, and stated it in qualified terms—as referring to major disturbances—because I believe that the potentiality of monetary policy in offsetting other forces making for instability is far more limited than is commonly believed." In his seminal work with Anna Schwartz, Friedman had laid the blame for the Great Depression on the inaction of the Federal Reserve. On Friedman's 90th birthday, then-governor of the Federal Reserve Ben Bernanke (2002) stated, "You're right, we did it. We're very sorry. But thanks to you, we won't do it again."

After becoming the Federal Reserve's chair in 2006, Bernanke was put to the test in 2008 as a financial crisis comparable to the one that triggered the Great Depression hit the US economy. At first, a new depression seemed imminent. But the Federal Reserve (and many other central banks) responded aggressively. By preventing bank failures, providing emergency credit to financial intermediaries, and increasing bank reserves, the central bank made sure that the money supply (as measured by M2) did not fall as precipitously as it did during the Great Depression; Friedman would have approved. At the same time, the Federal Reserve kept its focus on interest rates, now expanded through explicit forward guidance, and persistently increased the size of its balance sheet through quantitative easing policies that aimed to facilitate the operation of the mortgage market. This array of monetary policy actions arguably prevented a financial collapse and helped the economy recover (Blinder 2013). By the end, the US economic contraction lasted for 19 months and industrial output fell by 17 percent from peak to trough; during the Great Depression, the comparable numbers were 43 months and 52 percent. At least this one time, the Federal Reserve seems to have successfully rebutted Friedman's skepticism about its ability to respond to major disturbances.

The Road Ahead

Fifty years after Friedman's (1968) presidential address, it is remarkable that its themes remain central in the study of business cycles and monetary policy. Expectations, the long run, the Phillips curve, and the potential and limits of monetary policy all continue to be actively researched. Fifty years from now, our knowledge about each of these topics will surely be different, and we hope better, but we are willing to bet they will remain central topics in macroeconomics.

In the near future, the meager economic growth since the 2008–2009 recession may lead to a reexamination of Friedman's natural-rate hypothesis. At this point, the simplest explanation is that this stagnation is due to a slowdown in productivity unrelated to the business cycle. Alternatively, however, it might contradict Friedman's classical view of the long run, either because hysteresis effects set in after large recessions or because the economy can suffer from a chronic shortage of aggregate demand (as Blanchard discusses in this issue).

Either way, the Phillips curve has come a long way since A. W. Phillips first plotted the unemployment rate against the change in nominal wages using British data. As a scatter plot, the Phillips curve has shifted so often that no one takes it to be anything other than a transitory, reduced-form empirical relation. Yet as a synonym for nominal rigidities, in the sense of a structural two-way causal relation between nominal and real variables in the short run, the Phillips curve is as alive as ever. Much recent research has embraced Keynes's vision of focusing on how wages and prices are set at the micro level, both in theory and in the data. Future work might do well to re-embrace Friedman's vision and turn to modeling expectations instead for a better understanding of the Phillips curve (Mankiw and Reis 2002).

Focusing on expectations is especially promising in light of the active work in the area (Coibion, Gorodnichenko, and Kumar forthcoming). On the side of theory, researchers are using insights from behavioral economics about the ways people go about crafting their expectations together with the formalism provided by measures of limited information flows borrowed from computer science (Mankiw and Reis 2010; Sims 2010). On the side of data, there are a growing number of surveys on people's expectations, field experiments that show how news spreads in networks of people, and laboratory data on the formation of perceptions. The road ahead lies in combining these approaches to provide a better benchmark model of expectations that can replace both adaptive and rational expectations (Woodford 2013).

In addition, the role of monetary policy is in flux today and has drifted quite far from the topics that Friedman emphasized in his presidential address. The overall design of central banks does not just merit discussion, but is also the subject of revisions in practice (Reis 2013). The road ahead will likely lead to progress in four areas: the interaction between fiscal and monetary policy, the role of bank reserves, near-zero interest rates, and financial stability.

Friedman discussed fiscal policy in the presidential address only briefly by condemning the "cheap money policies after the war" for producing inflation in their futile attempt to keep interest payments on the debt low. Otherwise, fiscal

authorities are largely ignored in the address. Current research instead emphasizes that central banks cannot live in isolation from fiscal authorities. On the one hand, central banks are fiscal agents. Their choices have consequences for what the fiscal authorities can achieve and for the fiscal burden they face (Reis forthcoming). On the other hand, fiscal authorities affect financial stability through implicit guarantees that encourage risky behavior, can smooth or enhance the business cycle by alternating between stimulus and austerity, and can put pressure on inflation through unsustainable fiscal policy (Sims 2013). Discussions of monetary policy today often include their fiscal dimensions, even if briefly, but this was not the case in most of Friedman's address.

As central banks focus on interest rates and the use of currency declines, the old monetarism that emphasized the medium of exchange seems outdated. But in its place, a new monetarism is being built on the role of liquidity in financial markets and on the role that reserves play in these markets. This work builds on the fact that at the end of 2015, US banks held twice as much in reserves issued by the central bank as they did in government bonds issued by the Treasury (Reis 2016). Reserves are one of the largest homogeneous financial assets today, and the central bank can control both the interest it pays on them as well as their quantity independently. "Reservism" may become the new face of monetarism, not as a policy target but as an approach to inflation and as a guide for central banks for their "quantitative easing" policies and other uses of the central bank balance sheet (Benigno and Nisticò 2015).

Friedman had studied the Great Depression extensively, and his views on monetary policy were deeply influenced by this experience. It is therefore surprising that the challenges of near-zero interest rates receive scant attention in his presidential address. Implicitly, Friedman seemed to dismiss the Keynesian views that the power of monetary policy is compromised when interest rates are near-zero or that this requires the use of different monetary policy tools. Recent research on monetary policy takes a different perspective. It emphasizes that there is a lower bound on interest rates (slightly below zero) that puts a constraint on monetary policy, and suggests using forward guidance policies to overcome it, or raising the inflation target to reduce its occurrence (Eggertsson and Woodford 2003). Some go as far as to suggest radical changes to the monetary system, such as abolishing currency or introducing dual currencies, to deal with the constraint posed by the zero lower bound (Agarwal and Kimball 2015; Rogoff 2017). This research suggests that if real interest rates remain as low as they currently are for long, monetary policy in the future may look very different from the monetary policy that Friedman considered (Eggertsson and Mehrotra 2014).

Finally, Friedman was an expert on financial crises, yet in an address on monetary policy, he chose to ignore the interaction between monetary policy and financial stability. Of course, it had long been recognized that as the lender of last resort, the central bank has some responsibility for financial stability. Yet any desire to exert tight control over the level of asset prices is foolish for all the reasons that Friedman explained in his address, especially when applied to stock prices or house prices

(Brunnermeier and Schnabel 2016). Friedman would have been similarly skeptical about the current foray of central banks into macroprudential regulation (the use of financial regulatory tools to promote macroeconomic goals); the presidential address does not have a single word on regulation as a task for monetary policy. After almost a decade of research into financial crises, the current consensus in the literature seems to be that central banks should pay close attention to credit and funding variables in an attempt to forecast and prevent financial crises, should take into account the effect of their actions on financial intermediaries, and at times should use financial regulation to intervene directly when doing so would promote financial and macroeconomic stability (Adrian and Shin 2008; Brunnermeier and Sannikov 2013). There remain many hard questions about the role that central banks should play and about how much we should expect from these important institutions. But in the spirit of Milton Friedman's presidential address, we suspect that it would be best for central bankers to remain humble in what they aspire to achieve.

■ *We are grateful for comments from Andrea Alati, Charlie Bean, Denis Fedin, Mark Gertler, Robert Hetzel, Tina Liu, Maria Lopez-Uribe, Enrico Moretti, Edward Nelson, Timothy Taylor, and Nina Vendhan.*

References

Adrian, Tobias, and Hyun Song Shin. 2008. "Financial Intermediaries, Financial Stability, and Monetary Policy." In *Maintaining Stability in a Changing Financial System*, Jackson Hole Economic Policy Symposium, Federal Reserve Bank of Kansas City, 287–334.

Agarwal, Ruchil, and Miles S. Kimball. 2015. "Breaking Through the Zero Lower Bound." IMF Working Paper 15/224.

Andrade, Philippe, Richard K. Crump, Stefano Eusepi, and Emanuel Moench. 2016. "Fundamental Disagreement." *Journal of Monetary Economics* 83: 106–128.

Ball, Laurence. 1994. "Credible Disinflation with Staggered Price-Setting." *American Economic Review* 84(1): 282–89.

Ball, Laurence, N. Gregory Mankiw, and Ricardo Reis. 2005. "Monetary Policy for Inattentive Economies." *Journal of Monetary Economics* 52(4): 703–725.

Barro, Robert J. 1977. "Unanticipated Money Growth and Unemployment in the United States." *American Economic Review* 67(2): 101–15.

Benigno, Pierpaolo, and Salvatore Nisticò. 2015. "Non-Neutrality of Open Market Operations." CEPR Discussion Paper 10594.

Bernanke, Ben S. 2002. "On Milton Friedman's Ninetieth Birthday." Remark at the Conference to Honor Milton Friedman, University of Chicago, Chicago, Illinois. November 8. https://www.federalreserve.gov/BOARDDOCS/SPEECHES/2002/20021108.

Blanchard, Olivier. 2009. "The State of Macro." *Annual Review of Economics* 1(1): 209–28.

Blanchard, Olivier. 2016. "The Phillips Curve: Back to the '60s?" *American Economic Review* 106(5): 31–34.

Blinder, Alan S. 2013. *After the Music Stopped: The Financial Crisis, the Response, and the Work Ahead.* Penguin Press.

Blinder, Alan S., Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen.

2008. "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence." *Journal of Economic Literature* 46(4): 910–45.

Brayton, Flint, Andrew Levin, Ralph Lyon, and John C. Williams. 1997. "The Evolution of Macro Models at the Federal Reserve Board." *Carnegie-Rochester Conference Series on Public Policy* 47(1): 43–81.

Breit, William, and Barry T. Hirsch. 2004. *Lives of the Laureates: Eighteen Nobel Economists.* Fourth edition of *Lives of the Laureates.* MIT Press.

Brunnermeier, Markus K., and Yuliy Sannikov. 2013. "Redistributive Monetary Policy." In *The Changing Policy Landscape*, Jackson Hole Economic Policy Symposium, Federal Reserve Bank of Kansas City, 331–84.

Brunnermeier, Markus K, and Isabel Schnabel. 2016. "Bubbles and Central Banks: Historical Perspectives." Chap. 12 in *Central Banks at a Crossroads: What Can We Learn from History?* edited by Michael D. Bordo, Øyvind Eitrheim, Marc Flandreau, and Jan F. Qvigstad. Cambridge University Press.

Chari, V. V., and Patrick J. Kehoe. 2006. "Modern Macroeconomics in Practice: How Theory Is Shaping Policy." *Journal of Economic Perspectives* 20(4): 3–28.

Coibion, Olivier, and Yuriy Gorodnichenko. 2012. "What Can Survey Forecasts Tell Us about Information Rigidities?" *Journal of Political Economy* 120(1): 116–59.

Coibion, Olivier, and Yuriy Gorodnichenko. 2015. "Is the Phillips Curve Alive and Well After All? Inflation Expectations and the Missing Disinflation." *American Economic Journal: Macroeconomics* 7(1): 197–232.

Coibion, Olivier, Yuriy Gorodnichenko, and Rupal Kamdar. Forthcoming. "The Formation of Expectations, Inflation and the Phillips Curve." *Journal of Economic Literature.*

Eggertsson, Gauti B., and Neil R. Mehrotra. 2014. "A Model of Secular Stagnation." NBER Working Paper 20574.

Eggertsson, Gauti B., and Michael Woodford. 2003. "The Zero Bound on Interest Rates and Monetary Policy." *Brookings Papers on Economic Activity*, no. 1, p. 139–235.

Friedman, Milton. 1962. *Capitalism and Freedom.* University of Chicago Press.

Friedman, Milton. 1968. "The Role of Monetary Policy." Presidential address delivered at the 80th Annual Meeting of the American Economic Association. *American Economic Review* 58(1): 1–17.

Friedman, Milton. 1970. "The Social Responsibility of Business is to Increase Its Profits." *New York Times Magazine*, September 13.

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

Galí, Jordi, and Mark Gertler. 2007. "Macroeconomic Modeling for Monetary Policy Evaluation." *Journal of Economic Perspectives* 21(4): 25-46.

Goodfriend, Marvin, and Robert G. King. 1997. "The New Neoclassical Synthesis and the Role of Monetary Policy." *NBER Macroeconomics Annual* vol. 12 (1997), pp. 231–96.

Hall, Robert E. 1976. "Notes on the Current State of Empirical Macroeconomics." June. Hoover Institution Stanford University. Available at: https://web.stanford.edu/~rehall/All_publications.htm.

Hansen, Alvin H. 1953. *A Guide to Keynes.* McGraw-Hill.

Hetzel, Robert L. 2017. "What Remains of Milton Friedman's Monetarism?" Unpublished paper, FRB Richmond.

Hicks, J. R. 1937. "Mr. Keynes and the 'Classics'; A Suggested Interpretation." *Econometrica* 5(2): 147—59.

Keynes, John M. 1923. *A Tract on Monetary Reform.* Macmillan.

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

Kuznets, Simon. 1955. "Economic Growth and Income Inequality." *American Economic Review* 45(1): 1–28.

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

Lucas, Robert E. 1972. "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4(2): 103–124.

Malmendier, Ulrike, and Stefan Nagel. 2016. "Learning from Inflation Experiences." *Quarterly Journal of Econo*mics 131(1): 53–87.

Mankiw, N. Gregory, and Ricardo Reis. 2002. "Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *Quarterly Journal of Economics* 117(4): 1295–1328.

Mankiw, N. Gregory, and Ricardo Reis. 2010. "Imperfect Information and Aggregate Supply." Chap 5. in *Handbook of Monetary Economics,* vol. 3A, edited by Benjamin M. Friedman and Michael Woodford. Elsevier.

Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers. 2004. "Disagreement about Inflation Expectations." *NBER Macroeconomics Annual*, vol. 18 (2003), pp. 209–70.

Mankiw, N. Gregory, and David Romer. 1991. *New Keynesian Economics*, Vol. 1: Imperfect *Competition and Sticky Prices*, and Vol. 2: *Coordination Failures and Real Rigidities.* MIT Press.

Nelson, Edward. 2017. *Milton Friedman and Economic Debate in the United States, 1932–72.*

Unpublished book, University of Sydney.

Phelps, Edmund S. 1967. "Phillips Curves, Expectations of Inflation and Optimal Unemployment over Time." *Economica* 34(135): 254–81.

Phelps, Edmund S. 1968. "Money-Wage Dynamics and Labor-Market Equilibrium." *Journal of Political Economy* 76(4): 678–711.

Pierce, Andrew. 2008. "The Queen Asks Why No One Saw the Credit Crunch Coming." *Daily Telegraph*, November 5.

Reis, Ricardo. 2013. "Central Bank Design." *Journal of Economic Perspectives* 27(4): 17–44.

Reis, Ricardo. 2016. "Funding Quantitative Easing to Target Inflation." In *Designing Resilient Monetary Policy Frameworks for the Future*, Jackson Hole Economic Policy Symposium, Federal Reserve Bank of Kansas City, 423–478.

Reis, Ricardo. Forthcoming. "Can the Central Bank Alleviate Fiscal Burdens?" In *The Economics of Central Banking*, edited by David Mayes, Pierre Siklos, and Jan-Egbert Strum. Handbooks in Economics. Oxford University Press.

Rogoff, Kenneth S. 2017. *The Curse of Cash*. Princeton University Press.

Samuelson, Paul A., and Robert M. Solow. 1960. "Analytical Aspects of Anti-Inflation Policy." *American Economic Review* 50(2): 177–94.

Sargent, Thomas J. 2008. "Evolution and Intelligent Design." *American Economic Review* 98(1): 5–37.

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.

Schultz, Theodore W. 1961. "Investment in Human Capital." *American Economic Review* 51(1): 1–17.

Sims, Christopher A. 2010. "Rational Inattention and Monetary Economics." Chap. 4 in *Handbook of Monetary Economics*, vol. 3A, edited by Benjamin M. Friedman and Michael Woodford. Elsevier.

Sims, Christopher A. 2013. "Paper Money." *American Economic Review* 103(2): 563–84.

Smets, Frank, and Rafael Wouters. 2007. "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach." *American Economic Review* 97(3): 586–606.

Svensson, Lars E. O. 2010. "Inflation Targeting." Chap. 22 in *Handbook of Monetary Economics*, vol. 3B, edited by Benjamin M. Friedman and Michael Woodford, 1237–1302. Elsevier.

Taylor, John B. 1993. "Discretion versus Policy Rules in Practice." *Carnegie-Rochester Conference Series on Public Policy* 39(1): 195–214.

Taylor, John B. 2016. "The Staying Power of Staggered Wage and Price Setting Models in Macroeconomics." Chap. 25 in *Handbook of Macroeconomics*, edited by John B. Taylor and Harald Uhlig. Elsevier.

Woodford, Michael. 2003. *Interest and Prices: Foundation of a Theory of Monetary Policy*. Princeton University Press.

Woodford, Michael. 2007. "The Case for Forecast Targeting as a Monetary Policy Strategy." *Journal of Economic Perspectives* 21(4): 3–24.

Woodford, Michael. 2010. "Optimal Monetary Stabilization Policy." Chap. 14 in *Handbook of Monetary Economics*, vol. 3B, edited by Benjamin M. Friedman and Michael Woodford. Elsevier.

Woodford, Michael. 2013. "Macroeconomic Analysis without the Rational Expectations Hypothesis." *Annual Review of Economics* 5(1): 303–46.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 97–120

Should We Reject the Natural Rate Hypothesis?

Olivier Blanchard

Fifty years ago, Milton Friedman (1968) delivered his presidential address "The Role of Monetary Policy" at the December meetings of the American Economic Association and articulated what became known as the "natural rate hypothesis." It was a joint hypothesis, composed of two sub-hypotheses. The first was that there was a natural rate of unemployment, independent of monetary policy. To quote Friedman: "The 'natural rate of unemployment' . . . is the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on" (p. 8). The second was that monetary policy could not sustain unemployment below the natural rate without leading to higher and higher inflation, a proposition that became known as the "accelerationist hypothesis." Again, to quote Friedman: "There is always a temporary trade-off between inflation and unemployment; there is no permanent trade-off. The temporary trade-off comes not from inflation per se, but from unanticipated inflation, which generally means, from a rising rate of inflation. The widespread belief that there is a permanent trade-off is a sophisticated version of the confusion between 'high' and 'rising' that we all recognize in simpler forms. A rising rate of inflation may reduce unemployment, a high rate will not" (p. 11).

■ *Olivier Blanchard is C. Fred Bergsten Senior Fellow, Peterson Institute for International Economics, Washington, DC, and Robert M. Solow Professor of Economics Emeritus, Massachusetts Institute of Technology, Cambridge, Massachusetts.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.1.97

doi=10.1257/jep.32.1.97

For the sake of clarity, in the rest of the paper, I shall refer to the joint hypothesis as the "natural rate hypothesis," and the two separate sub-hypotheses as the "independence hypothesis" and the "accelerationist hypothesis." Notice that they are separate hypotheses. The implications will be different if either one fails separately, or if both fail. Notice also that, while Friedman referred to unemployment, he clearly had in mind output more generally. The natural rate hypothesis can be recast in terms of output: that is, potential output is independent of monetary policy, and there cannot be sustained deviations of output above potential without increasing inflation. Thus, in this paper, I shall look at the evidence for both unemployment and for output.

Together, the two hypotheses have very strong implications. If inflation is to remain stable, periods during which output exceeds potential output must be offset by periods during which output is below potential; in other words, booms must be fully offset by slumps. Monetary policy cannot do more, and indeed should not try to do more, than smooth fluctuations around the independent path of potential output.

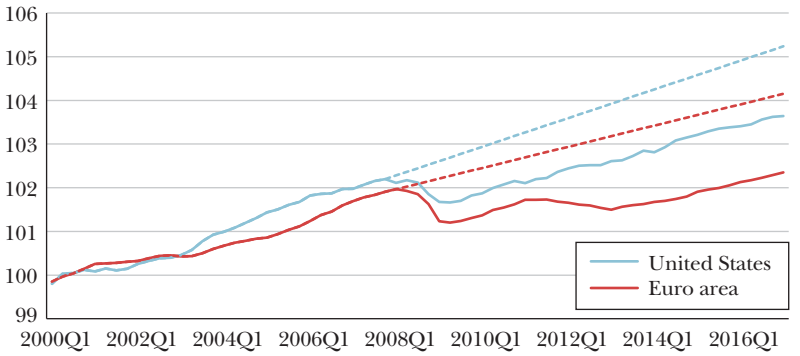
While the natural rate hypothesis was controversial at the time, it quickly became widely accepted, and has been the dominant paradigm in macroeconomics ever since. It is embodied in the thinking and the models used by central banks, and it is the basis of the inflation-targeting framework used by most central banks today.

However, there have always been grumblings about the extent to which this hypothesis fully characterizes the world, about whether potential output is really independent of monetary policy, and about whether there really is no long-run trade-off between inflation and output. In the 1980s in particular, the natural rate of unemployment in Europe appeared to increase following every recession, and the idea that high actual unemployment might cause an increase in the natural rate became more popular. In Blanchard and Summers (1986), Larry Summers and I argued that hysteresis, which refers to the theory that changes in the natural rate of unemployment can be path-dependent (an idea which could be traced at least to Phelps 1972), could be the explanation for this increase. Then, over time, as the so-called Great Moderation took place from the mid-1980s up to about 2007, research on hysteresis largely disappeared.

But recently grumblings have increased (for example, Cœuré 2017). This is for two reasons, both linked to the Great Financial Crisis and the accompanying recession. First, the level of output appears to have permanently been affected by the crisis and its associated recession. This is shown in Figure 1, which plots the evolution of (the log of) GDP since 2000 for both the United States and the European Union, both normalized to equal 100 in 2000. In both cases, it appears as if the output path has shifted down and is now increasing along a lower trend line than before the crisis. This pattern led Summers (2014) to state: "Any reasonable reader of the data has to recognize that this financial crisis has confirmed the doctrine of hysteresis more strongly than anyone could have anticipated."

Second, in contrast to the accelerationist hypothesis, very high unemployment did not lead to lower and lower inflation, but rather just to ongoing low inflation.

Figure 1
Advanced Economies (log) Real GDP and Extrapolated Trend
(index equals 100 in 2000)



Source: Data from the US Bureau of Economic Analysis and the Statistical Office of the European Communities.
Note: The figure shows the log real GDP since 2000 for both the United States and the European Union, normalized to equal 100 in 2000. The log linear trend is estimated over 2000Q1 to 2007Q4, and extrapolated up to 2017Q1.

In both the United States and the European Union, except for the large decline in inflation in 2009, there does not appear to be any relationship between the unemployment rate and the change in inflation in the last two decades. We appear to have returned instead to a relation between the unemployment rate and the rate of inflation, rather than between the unemployment rate and the change in the rate of inflation.

Neither fact is by itself a clear rejection of the natural rate hypothesis. It could be that the decrease in output relative to trend reflects a decrease in the underlying trend, or strong and persistent effects of the financial crisis on the supply side of the economy, rather than adverse, hysteretic effects of lower output perpetuating itself. If so, the outcome of the Great Financial Crisis might carry no implication for the effects of monetary policy shocks. Moreover, the Lucas critique of the Phillips curve has told us that expectations matter, and an apparent trade-off between unemployment and inflation may well disappear when circumstances change or when the policymaker tries to exploit it. Yet, these facts, and the 50th anniversary of Friedman's (1968) AEA presidential address, suggest that it is a good time to review the available evidence.

The paper assesses what we know and do not know. I begin by revisiting the logic of the independence hypothesis and looking at the macroeconomic and microeconomic evidence. I then turn to the evidence on the accelerationist hypothesis. Finally, I consider potential policy implications and conclude. To anticipate the answer to the question in the title: I see the macroeconomic and the microeconomic evidence as suggestive but not conclusive evidence against the natural rate hypothesis. Policymakers should keep the natural rate hypothesis

as their null hypothesis, but also keep an open mind and put some weight on the alternatives.

On the Independence Hypothesis: Persistence versus Permanence

The first step must be to recast the discussion. The discussion about the independence hypothesis has largely taken the form of a choice between what appear to be two sharply different classes of models: "standard models" where monetary policy does not affect potential output, and "hysteresis models," where monetary policy has permanent effects on potential output. The seeming dichotomy between these models is misleading. Even in the most standard models, monetary policy is likely to affect potential output for some time. Conversely, in most hysteresis models, the effects of monetary policy are likely to be persistent, but not necessarily permanent. The issue is thus about the size and persistence of the effects of monetary policy on potential output, not their existence nor their permanence.

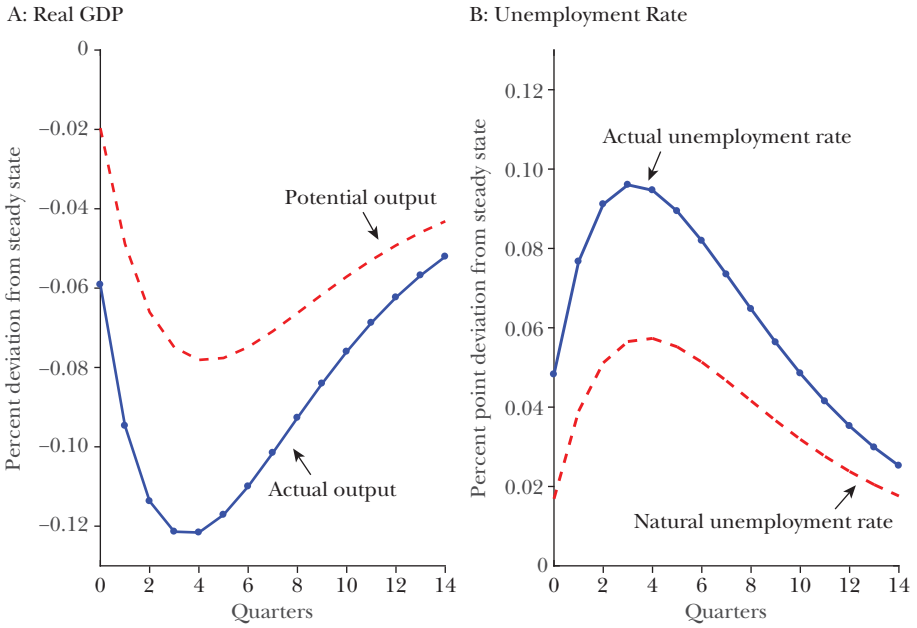
Let me spell out these points more precisely. The discussion must start with a definition of "potential output." I define potential output as the level of output which would obtain if, given actual history, nominal price and wage rigidities were lifted from now on. I define the natural rate of unemployment in the same way. Potential output is sometimes defined as the level of output that would obtain if nominal rigidities had always been absent, both in the past and in the future. Potential output, defined this way, would be tautologically independent of monetary policy. But this is not a useful definition, because what matters is what output can be today, as opposed to what it could have been in some hypothetical world.

Now take any "standard model" in which, because of nominal rigidities, monetary policy does affect output for some time. Suppose that tighter monetary policy triggers a recession and a decline in output. This decline in output is likely to come with a decline in investment. Thus as output declines, the capital stock is lower for some time, and by implication, so is potential output. The same may be true of other factors of production. For example, if matching frictions prevent employment from quickly returning to its pre-recession level, then capital and labor are quasi-fixed in the short run. Thus, potential output, the output that would prevail if all nominal rigidities were suddenly lifted, may be fairly close to actual output, and be affected for some time by monetary policy.

As an example, Figure 2 shows the behavior of actual and potential output, as well as the behavior of the actual and the natural unemployment rate, in a model by Christiano, Eichenbaum, and Trabandt (2016). This model allows for capital accumulation and also for matching frictions in the labor market, frictions that prevent unemployment from quickly returning to its long-run natural rate. After an adverse monetary policy shock, potential output follows a path similar in shape to that of actual output, but with an amplitude of about half. When actual output reaches its trough of −0.12 percent, potential output also reaches its trough, at about −0.07

Figure 2

Impulse Responses to a Contractionary Monetary Policy Shock (50 Annual Basis Points)



Source: The figure comes courtesy of a simulation by Mathias Trabandt.
Note: Figure 2 shows the behavior of actual and potential output, as well as the behavior of the actual and the natural unemployment rate, in a model by Christiano et al. (2016). Potential output in each quarter t is derived as the level of output which would prevail in quarter t if, given history up to quarter t, all nominal rigidities were removed from quarter t on. The natural rate of unemployment in each quarter is defined similarly.

percent. Fifteen quarters out, potential output is still −0.04 percent lower than before the shock. In other words, even if all nominal rigidities were removed at that point, output would still be −0.04 percent lower than absent the monetary policy shock. A similar pattern holds for the natural and actual unemployment rates.[1]

Now consider the channels that have been emphasized in hysteresis models. Some of these channels may indeed imply a permanent effect of monetary policy. For example, if a recession leads to lower research and development for some time, and if total factor productivity depends in part on the accumulation of past research and development efforts, then total factor productivity may indeed be permanently lower than it would have been, absent the recession. But some of the channels

[1] The magnitudes are small, for two reasons: the monetary shock is small, and the effects of monetary policy on actual output in the model are small as well. However, these aspects are not relevant to the point made in the text. I am thankful to Mathias Trabandt for performing these simulations. The original paper does not show the path of potential output or the path of the natural rate.

that have been studied suggest persistent effects rather than permanent ones. For example, if some of the long-term unemployed become unemployable, the effect will eventually disappear as these workers reach the age at which they would have stopped working anyway.

In short, all relevant models imply an effect of monetary policy on potential output and on the natural rate that will last for some time.[2] The goal of the empirical work must be, at the macro level, to assess the degree of persistence of the effects, and, at the micro level, to identify and examine specific channels of persistence.

Macro Evidence on the Independence Hypothesis: Monetary Policy, Recessions, Unemployment, and Output

The independence hypothesis is about the effects of monetary policy shocks on potential output. Thus, the first issue is how to identify monetary policy shocks. One approach would be to use a vector autoregression (VAR) methodology with identified monetary policy shocks, trace their dynamic effects on output or unemployment, and assume that, as the horizon increases, these increasingly reflect their effects on potential output and the natural unemployment rate. Given the well-known difficulties of identifying those monetary shocks, and the statistical uncertainty associated with impulse responses, this approach does not look very promising. A meta-study of vector autogression studies by De Grauwe and Costa Storti (2004) finds the mean effect of a 1 percent interest rate shock on output to be −0.15 percent after five years. However, the distribution of estimates has a standard deviation of 0.27 percent, so a zero effect is not far from the middle of the distribution.

A more promising approach is to look at recessions associated with intentional disinflations. The shocks are clearly monetary shocks; they are large; and they are plausibly largely exogenous, reflecting more a change in policy rather than the response of policy to other shocks. This approach has been pursued by a number of authors, in particular Laurence Ball in a number of contributions (for example, Ball 2009, 2014). It is the approach I shall follow here, building on Blanchard, Cerutti, and Summers (2015). Details are given in the online appendix, but the general approach is as follows.

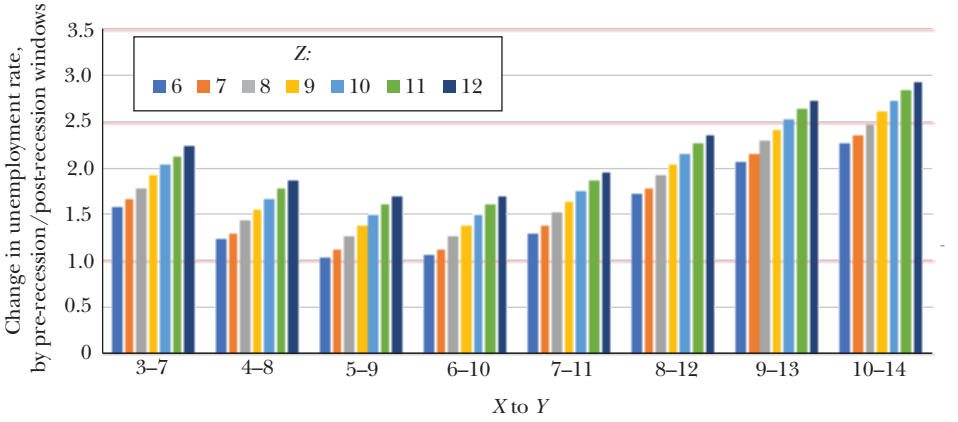
I consider 22 advanced economies over the last 50 years, and, using a simple algorithm looking at peaks and troughs (based on Harding and Pagan 2002), I identify 122 recessions. I then classify recessions according to their proximate cause—such as intentional disinflations, oil price increases, financial crises, and so

[2] Put more formally, all realistic models will have a number of state variables, some of them affecting potential output. Thus, even if all nominal rigidities were suddenly removed during a recession, potential output would be different from what it would have been absent the shock. One can think of hysteresis models as models focusing on those mechanisms potentially leading to long-lasting, or in the limit, permanent effects on some of these state variables, and by implication on output and unemployment.

Figure 3

Change in Unemployment Rate after Disinflation Recessions

(Average unemployment rate X to Y years after the recession – Average unemployment rate 2 to Z years before the recession)



Source: Author's calculations.

Note: I consider 22 advanced economies over the last 50 years, identify 122 recessions, and focus on the 22 recessions associated with intentional disinflations. There is a trade-off in looking at averages over different time intervals. The longer the length of the post-recession period for example, the more the average can tell us about persistence, but the more the average is affected by other shocks than the disinflation. Each bar shows (Average unemployment rate *X* to *Y* years after the recession) minus (Average unemployment rate 2 to *Z* years before the recession). Each set of bars corresponds to a given post-recession average but different pre-recession averages. Thus, the first set gives results for the post-recession average computed over 3 to 7 years after the end of the recession, and pre-recession averages computed over 2 to 6 years before the beginning of the recession for the first bar, 2 to 7 years for the second bar and so on. (In all cases, I leave out the two years before the recession in case there was a cyclical boom.)

on—and focus first on the 22 recessions associated with intentional disinflations. (I shall return to whether and how one can use information from the other recessions later.) I then compute the average unemployment rate over various time intervals both pre- and post-recession, and take the difference between the post- and pre-recession periods. As discussed in Blanchard, Cerutti, and Summers (2015), there is a trade-off in looking at averages over different time intervals. The longer the length of the post-recession period for example, the more the average can tell us about persistence, but the more the average is affected by other shocks than the disinflation.

The results are shown in Figure 3. The different bars correspond to the different time intervals used to compute pre-recession and post-recession averages. Each set of bars corresponds to a given post-recession average, but different pre-recession averages. Thus, the first set gives results for the post-recession average computed over 3 to 7 years after the end of the recession, with pre-recession averages computed over 2 to 6 years before the beginning of the recession for the first

bar, 2 to 7 years for the second bar, and so on. (In all cases, I leave out the two years before the recession in case there was a cyclical boom.) The second set of bars gives the results for the post-recession average computed over 4 to 8 years, and so on. The visual impression is fairly clear, with large and very persistent increases in unemployment on average after those recessions, with differences ranging from 1 to 3 percent depending on the combination.[3]

Three caveats are in order. First, the majority of recessions in this category took place around the same time in the early 1980s, so the results may reflect common time effects to some extent. Second, the averages hide some heterogeneity. For the combination of time intervals that gives the smallest increase, only 15 out of the 22 recessions are associated with increased unemployment; for the combination of time intervals which gives the largest increase, the number increases to 19. Third, the figure shows the changes in the actual unemployment rate, not necessarily the changes in the natural unemployment rate. While it is plausible that the two may converge, as we look at longer and longer intervals pre- and post-recession, one may worry that this is not the case. This is where, in principle, the behavior of inflation can offer more information. During the 1980s and early 1990s when these recessions took place, the evidence (reviewed later in this paper) is that the accelerationist Phillips curve gave a good characterization of the data, so we can look at the change in inflation as a signal of the distance between actual and natural unemployment rates. The average annual change in the inflation rate over the various pre-recession time intervals ranges from 0.04 to 0.12 percent. The average annual change over the various post-recession intervals ranges from −0.40 to 0.12 percent. These numbers are small and suggest that the change in the actual unemployment rate can be interpreted mostly as a change in the natural rate.

While Friedman's (1968) natural rate hypothesis focused on unemployment, along with much of the research on hysteresis, his arguments clearly were meant to apply to output as well. Thus I use a similar methodology to look at whether output returns to its pre-recession level after recessions triggered by intentional disinflations. More specifically, I estimate a log-linear trend for output over some pre-recession time interval, extrapolate it post-recession, and compute the resulting output gap as the average difference between actual and extrapolated output over some post-recession time interval. One delicate empirical issue is that output growth has declined in most advanced countries over the sample period; thus, the extrapolation of a log-linear trend over any pre-recession time interval will tend to overpredict post-recession output and lead to an estimated negative output gap, even in the absence of any hysteresis. I correct for this decrease in the underlying trend when extrapolating the pre-recession trend using a method described in the online Appendix.

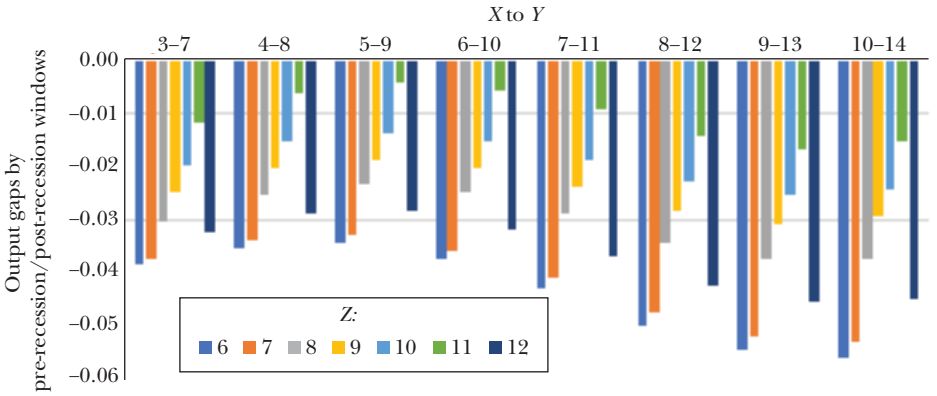
The results are shown in Figure 4 for the various pre-recession and post-recession time intervals. The average output gaps are typically negative, but the results

[3] As shown in the online Appendix, the results are very similar if the prime-age male unemployment rate is used instead of the overall unemployment rate.

Figure 4

Output Gaps after Disinflation Recessions

(Average output gap X to Y years after the recession based on extrapolated trend estimated using 2 to Z years before the recession)



Source: Author's calculations.

Note: For 22 recessions associated with intentional disinflations (see note to Figure 3), I consider average output gaps *X* to *Y* years after the recession relative to an extrapolated trend estimated 2 to *Z* years after the recession.

are less consistent than for unemployment, and are more sensitive to the time intervals used to estimate the pre-recession gap. Also, the averages hide substantial heterogeneity. While the average gaps are negative, the proportion of negative output gaps over all time intervals and all episodes is only 55 percent, which offers only weak support for the hysteresis hypothesis.

An obvious question is how to reconcile the unemployment and the output results. To make progress, I decompose the log output gap between a log employment gap and a log productivity gap, using for each the same methodology as for log output, so the sum of the two gaps is equal to the output gap. The results (reported in the online Appendix) lead to three main conclusions: The employment gaps are consistently negative, largely insensitive to the choice of time interval, and close to the unemployment gaps reported above. The productivity gaps are, perhaps surprisingly, often positive, and are sensitive to the choice of pre-recession time interval. A tentative explanation for this sensitivity is, again, that most disinflation episodes took place around the same time, and one unusual pre-recession year can affect the results quite strongly.

Can the evidence from the other 100 recessions, those not caused by explicit disinflation efforts, be used to learn about persistent effects of monetary policy shocks? The answer is yes, but only if one is willing to make further assumptions. If, for example, one assumes that long-run labor supply is inelastic—be it individual labor supply or the relation between the wage and unemployment derived from matching-bargaining models—then if one finds persistent effects of nonmonetary

shocks on unemployment, this suggests the presence of channels relevant for monetary shocks also.

As an example, consider the case of recessions brought on by oil price increases. To the extent that such shocks are persistent, they imply a decrease in real consumption wages relative to trend; but if long-run labor supply is inelastic, this should eventually have no effect on unemployment. What about output? Theory suggests that, while oil shocks should not have a direct effect on productivity (as productivity is the ratio of value added—which, if correctly measured, is unaffected—to employment), they may have an indirect effect on productivity growth. For example, to the extent that technological progress is directed by shifts in economic incentives, a change in oil prices may well lead to a temporary slowdown in productivity growth as firms have to explore technologies corresponding to the new configuration of relative prices. If so, the productivity level may be lower in the long run than it would have been absent the increase in the oil price. Or take the case of recessions brought on by financial crises. By the same argument, if long-run labor supply is inelastic, one would not expect them to lead to a permanent increase in unemployment. But to the extent that recessions have persistent effects on financial intermediation, be it because of changes in behavior or changes in regulation, they may well also have persistent adverse effects on productivity. For example, banks may become more risk averse, financing projects with lower risk but also, by implication, a lower expected rate of return.

This suggests we should focus on unemployment rates using the same methodology as I used earlier for recessions associated with intentional disinflations. Looking at the 33 oil-related recessions (all of them taking place from the mid-1970s to the early 1980s) or the 19 recessions brought on by financial crises (12 of them taking place in the late 2000s), the evidence in both cases is of large, highly persistent, increases in unemployment, consistent across pre-recession and post-recession time intervals.[4] However, the same caveats apply as for the disinflation-triggered recessions before: In particular, the oil-price-related recessions all happen around the two episodes of large oil price increases in the mid and late 1970s, and thus the results could reflect common time effects. The same is true for the majority of financial-crisis–related recessions. Nevertheless, the fact that most recessions are associated with a positive unemployment gap is quite striking.

A similar exercise using recessions caused by nonmonetary shocks but focusing on output cannot be used to test the independence hypothesis, for the reasons discussed above: the results might be specific to the type of shock and not be relevant for monetary shocks. These results are still worth reporting briefly. Output gaps associated with oil-price-related recessions are negative, large, and consistent across time intervals. In contrast to the disinflation-related recessions, they reflect mostly productivity gaps. Output gaps associated with financial-crisis-related recessions are smaller, but also consistent across time intervals. They reflect mostly employment

[4]The method of classification and the detailed results, both for effects on employment and for effects on output, are described in the online Appendix.

gaps rather than productivity gaps. The current case of the United States stands as an exception: output remains far below its pre-crisis trend, as shown in Figure 1, but the unemployment rate is back to its pre-crisis level. It also shows the limits of the method I have used. There is fairly wide agreement that, at least in the US economy, the productivity growth slowdown started, in fact, a few years before the crisis (for example, Fernald 2015). If so, the methodology I have used attributes this decrease in productivity (and by implication the decrease in output) incorrectly to the recession.

To summarize: I read the macroeconomic evidence as suggestive of persistent effects of monetary policy on the natural unemployment rate and potential output. But the evidence is not overwhelming. Moreover, looking just at recessions has its limits: It cannot answer whether there are symmetrical effects of booms and recessions, which is a crucial issue for the design of policy. In this context, a closer look at potential channels of persistence and more microeconomic evidence may help to assess potential nonlinearities or asymmetries between recessions and booms.

Micro Evidence on the Independence Hypothesis: Channels for High Persistence

Persistent effects of monetary shocks on output may come either from employment or from productivity. Starting with employment, the initial mechanism emphasized in Blanchard and Summers (1986) focused on wage formation. In its simplest form, the argument was straightforward. Suppose that employed workers (or the unions representing them) set wages and did not care about the unemployed. Unemployment would then play no role in wage setting and would follow a random walk with no tendency to return to any particular value. After an adverse shock and a recession, it would remain higher. After a boom, it would remain lower. A modern treatment along these lines, in a micro-founded New-Keynesian model with insiders and outsiders, is given by Gali (2016) and shows the long-term effects of monetary shocks.

Our earlier argument correctly emphasized the power of insiders in wage formation, but it was too strong. Even if the employed workers do not care about the unemployed, they should care about their own situation, were they to become unemployed. The higher the unemployment rate, the more willing they should be to accept a lower wage. Also, wages are not set unilaterally by workers (or by unions), but rather unilaterally by firms or by a process of bargaining between firms and workers. In this case, wages will reflect the option of firms to hire the unemployed. The higher the unemployment rate, the larger the pool of potential hires, the stronger the firms will be in bargaining. For both reasons, even with selfish insiders, unemployment will matter.

One of the major research developments of the 1980s and 1990s was the development of a framework capturing these aspects, based on matching and bargaining, with the basic framework now known as the DMP model, for the work by Diamond,

Mortensen, and Pissarides. It gives a better way to think about the effect of unemployment on wages, and how the strength of the effect depends on the structure of the labor market and on labor market institutions.[5] For example, consider the potential role of employment protection. The higher the firing cost, the smaller the risk for an employed worker to become unemployed (leaving aside the risk of bankruptcy and firm closure) and the smaller the effect of unemployment on the wage. The higher the hiring cost, the smaller the risk for an employed worker of being replaced by an unemployed worker, and thus the smaller the effect of unemployment on the wage. In the limit, with high hiring and firing costs, unemployment may indeed have little effect on the wage and lead to highly persistent effects of monetary policy shocks on the natural rate of unemployment. This analytical framework suggests that high persistence is more likely in countries with high employment protection, more generous unemployment benefits, and stronger unions. An in-depth analysis, both theoretical and empirical, of the effect of such cross-country differences on the persistence of shocks on the natural rate and potential output remains, however, mostly to be done.[6]

A subsequent explanation for hysteresis focused on the effect of high unemployment on labor market institutions, and by implication, on the natural unemployment rate. Indeed, the high unemployment rate triggered by the two oil shocks of the 1970s led to an increase in unemployment protection and in the generosity of unemployment benefits in most European countries (Blanchard and Wolfers 2000). While these measures were taken to limit the initial increase in unemployment and make it less painful, it is likely they increased the natural rate. However, this explanation is specific to those recessions, and does not provide for a general channel of high persistence.

Yet another channel for hysteresis, and at this point probably the most popular one among researchers, has focused on the effect of high unemployment on the morale, skills, and employability of the long-term unemployed. It has long been known that the probability of becoming employed decreases with the duration of unemployment. For example, based on data from the Current Population Survey for 1994–2016, the average probability of becoming employed in the following month decreases from 28 percent if unemployed for less than 27 weeks to 14 percent if unemployed for more than 27 weeks. At the end of 2009, when the US unemployment rate reached a high of 10 percent, the probability of re-employment in the following month was 18 percent if unemployed for less than 27 weeks, but only 10 percent if unemployed for more than 27 weeks.

While these comparisons are suggestive, they do not prove that the long-term unemployed become less employable. It may be instead that the workers who are

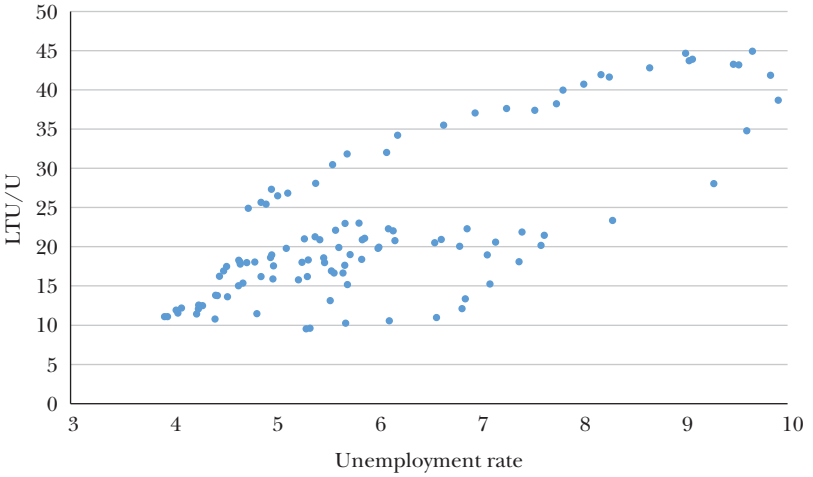
[5] Indeed, the model by Christiano, Eichenbaum, and Trabant used to generate Figure 2 above incorporates a formalization of the labor market reflecting matching and bargaining. Unemployment is a state variable, leading to a persistent, but not permanent, effect of monetary policy shocks on the natural rate.
[6] In Blanchard and Wolfers (2000), we took a first pass at it, by looking at the interaction of shocks and institutions in determining effects of shocks on unemployment. But much remains to be done.

the most employable are hired first, and thus, the longer the duration, the less employable is the remaining pool. However, two recent papers suggest that hysteresis—in this case, the effect of unemployment duration on employability—might be at work. First, Krueger, Kramer, and Cho (2014) use the time structure of the Current Population Survey (in the sample for four months, out for eight months, back in for four months) to look at the more relevant longer transition probabilities, and confirm the message from monthly probabilities. On average, for the period 1994 to 2012, the average probability of being employed 15 months later was 55 percent for those unemployed for less than 27 weeks, but only 40 percent for those unemployed for more than 27 weeks. Second, Abraham, Haltiwanger, Sandusky, and Spletzer (2016) link data from the Current Population Survey and the unemployment benefit register, and look at the employment history of the long-term unemployed. They find that the probability of being employed eight quarters earlier is roughly similar for the short-term and the long-term unemployed. If we think of this probability as a proxy for workers' characteristics, this suggests that, at the start of their unemployment spell, the long-term unemployed have roughly the same characteristics as the short-term unemployed, and that their lower probability of becoming employed is primarily caused by the duration of their unemployment rather than by their unobservable characteristics.

One more step is needed, however, to prove the case for hysteresis. It could be that the decreased probability reflects mostly the fact that firms, when they have the choice, often give priority in hiring to those who have been unemployed the least time—a decision rule that Peter Diamond and I (Blanchard and Diamond 1994) have called "ranking." If this is the case, so long as unemployment is high and firms get many applicants, the long-term unemployed will be less likely to get a job. But as unemployment decreases and the number of job applicants declines, the long-term unemployed will be more often at the front of the line, and see their relative probability of employment increase. What might appear like hysteresis in the short term will fade over time. While this hypothesis can be formally tested by looking at relative probabilities of employment as a function of overall unemployment, I have not seen it done. The regressions of transition probabilities for short-term and long-term unemployed on the overall unemployment rate by Krueger, Cramer, and Cho (2014, see their table 2) indicate that the relative probabilities do not vary much with the state of the labor market. If so, the data can be seen as providing some evidence for hysteresis.

To the extent that decreased employability is a source of hysteresis, one can then explore nonlinearities and asymmetry between recessions and booms. As shown in Figure 5, leaving aside short-run dynamics (which lead to countercyclical loops), the ratio of long-term unemployment to total unemployment in the United States is strongly increasing in unemployment. Put another way, the long-term unemployment rate is convex in the unemployment rate. If we think of the number of workers who become unemployable as roughly proportional to the number of long-term unemployed, this implies that hysteresis is asymmetric, being more relevant in recessions than in booms.

Figure 5
Ratio of Long-Term Unemployment (LTU) to Total Unemployment (U), against the Unemployment Rate, United States, 1990Q1–2016Q4

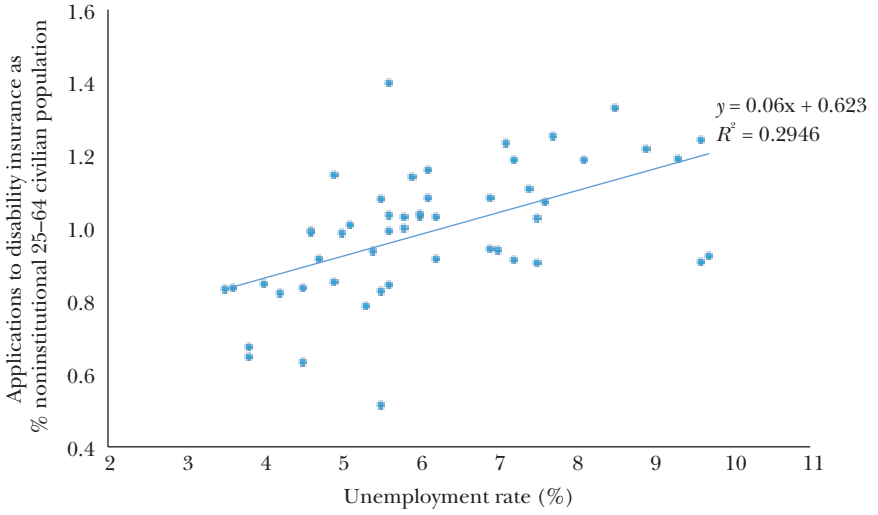


Source: Author using data from the Bureau of Labor Statistics.

If some workers become less employable or become discouraged, then the unemployment statistics will fail to capture hysteresis effects fully, because many of these workers will drop out of the labor force. Indeed, a recent question has been whether, in the United States, the high unemployment due to the financial crisis has contributed to the drop in labor market participation from 66 percent in 2007 to 63 percent at the end of 2016. The question is difficult to answer, because there has been a downward trend in the labor force participation rate since about 2000 due to the demography of an aging population. Aaronson et al. (2014) conclude that much if not most of the recent evolution of participation can be explained by the trend rather than by the crisis. However, a careful study by Yagan (2017) reaches a different conclusion. Yagan looks at the employment status of individuals in 2015 as a function of the increase in the unemployment rate in their local market from 2007 to 2009 controlling for individual characteristics, and he concludes that a 1 percent increase in the local unemployment rate in 2007–2009 led to a 0.4 percent decrease in the probability of being employed in 2015. His estimates imply that of the 7.2 percent decrease in the employment rate from 2007 to 2015 of the birth cohorts aged 30–49 in 2007, 4.8 percent can be attributed to demographics and 1.8 percent can be explained by the hysteretic effect of high unemployment during the Great Recession.

A complementary approach is to look at the evidence on disability insurance. Evidence on both applications and acceptances is useful. Cyclical variations in applications for disability insurance can give information about the loss of morale among workers as a result of the state of the labor market. And once people are accepted and start receiving disability payments, terminations are rare, except for infrequent

Figure 6
Applications to Disability Insurance versus Unemployment in the United States, 1960–2014



Note: Figure 6 plots, on the y-axis, applications to disability insurance in the United States as proportion of the 25–64 year-old population, and on the x-axis, the unemployment rate each year for the period 1965 to 2014.

program clampdowns (Autor and Duggan 2006). This implies that, to the extent that recessions lead to increases in disability insurance rolls, they have a hysteretic effect on the labor force.

Figure 6 plots, on the y-axis, applications to disability insurance in the United States as a proportion of the 25–64 year-old population, and on the x-axis, the unemployment rate each year for the period 1965 to 2014. The relation is strong, and both statistically and economically significant. An increase in the unemployment rate from say 5 to 10 percent increases the disability application ratio by 0.3 percent (or about 600,000 workers). If one takes the sum of annual unemployment rates in excess of 5 percent since 2008, which is roughly equal to 20 percent, this implies an additional 2.4 million more disability applications, and given an acceptance rate of about 35 percent, a permanent reduction in the labor force of about 800,000 workers. This channel may be seen as a strong piece of micro evidence in favor of hysteresis, relevant not just for disability insurance, but for the effect of unemployment on labor supply more generally. (In contrast to the previous graph, however, there is no evidence of a convex relation between applications and unemployment, thus no evidence of asymmetry between the effects of high and low unemployment.)

The macroeconomic evidence given earlier, suggested that, at least for disinflation-related recessions, the main channel of persistence was through employment

rather than through productivity. Nevertheless, it is useful to briefly explore this second potential channel as well.

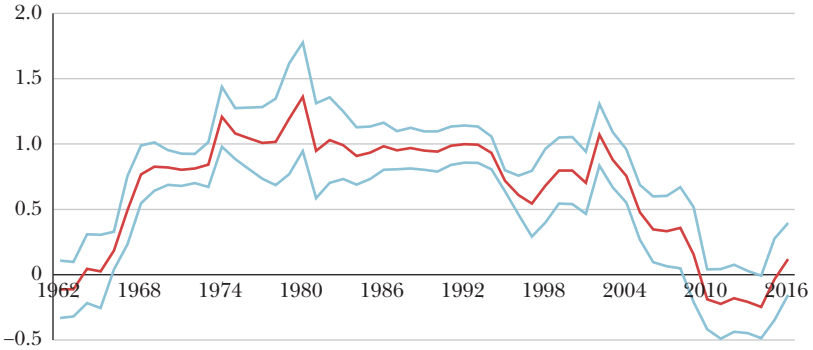
I discussed earlier the role of lower capital accumulation in leading, during the recession, to a decrease in labor productivity given total factor productivity. Rough computations suggest that the decline in the capital stock during a typical recession, and by implication the effect on labor productivity given total factor productivity, is small. However, theory suggests that recessions could have a permanent effect on total factor productivity itself and, by implication, on labor productivity. If we think, somewhat simplistically, of total factor productivity as being determined in part by the sum of past spending on research and development, then lower research and development during a recession will lead to permanently lower total factor productivity (and a boom will do the reverse). However, the empirical evidence suggests again limited effects: A regression of the rate of change of research and development spending on the rate of change of GDP for the period 1960–2013 for the United States delivers a low R^2 and a coefficient of about 1. This coefficient implies that a 1 percent decrease in GDP is associated with a decrease in research and development spending of 1 percent—a small effect.

Another potential way in which recessions may affect total factor productivity is through their effect on the speed of adoption of inventions. Anzaotegui, Comin, Gertler, and Martinez (2016) look at the effects of (detrended) GDP per person on the speed of adoption of 26 technologies in the United States and the United Kingdom over the period 1947 to 2003. They find that low activity indeed has a negative effect on the speed of adoption. However, to the extent that full adoption still eventually takes place, this suggests only a temporary slowdown in productivity growth—and persistence rather than permanence of the effects of recessions.

Yet another channel discussed in the literature is the effect of recessions on reallocation, and in turn on productivity growth. The sign of the effect is a priori ambiguous. Recessions may force low-productivity firms to close sooner, leading to more creative destruction and an increase in productivity. Alternatively, if the bankruptcy process is inefficient, it may instead force some high-productivity firms with high debt to close, leading to a decrease in productivity (for example, see Caballero and Hammour 1994). Looking at past US recessions, Foster, Grim, and Haltiwanger (2014) estimate that the effect has been generally positive and surprisingly large. Other things equal, and with the exception of the recession associated with the financial crisis, the reallocation due to recessions has typically led to increases in productivity growth of 0.4 to 0.8 percent depending on the depth of the recession (a result which fits the finding of often positive productivity gaps in disinflation-induced recessions in the previous section).

To summarize: of the microeconomic channels potentially behind high persistence, the most persuasive one appears to be that high unemployment leads some workers to be less employable or to give up on looking for jobs, increasing unemployment or reducing the labor force, and by implication, leading to a persistent effect on potential output.

Figure 7
The Evolving Phillips Curve: The Evolution of the Coefficient on Lagged Inflation in a Regression of Inflation on the Unemployment Rate
(*with +/– 1 standard deviation)*



Source: Bureau of Labor Statistics.
Note: The figure gives the evolution of the coefficient on lagged inflation in a simple specification regressing inflation as measured by the rate of change of the Consumer Price Index on a constant term, itself lagged, and the unemployment rate, using annual data since 1948, and backward-looking rolling samples of 15 years. It shows the increase in the coefficient early on, the long period of stability around 1, and the sharp recent decline. The coefficient today is not significantly different from zero.

The Accelerationist Hypothesis

The story of the changing Phillips curve—the relation between inflation and unemployment—has been told many times. Soon after Friedman's (1968) presidential address, and just as he had predicted, the trade-off between the unemployment rate and the inflation rate that had characterized the 1960s started to weaken, replaced in time by the "accelerationist Phillips curve," a relation between the unemployment rate and the change in the inflation rate. Put another way, the coefficient on lagged inflation in the Phillips curve steadily increased from a value close to 0 to a value close to 1.

This shift was documented in real time during the 1970s. For example, Perry (1970) estimated the coefficient on lagged inflation to be 0.34, while Perry (1978) estimated the same coefficient to be 1.0. Starting in the 2000s, however, the coefficient has sharply declined, and appears now to be again close to zero. This is shown in Figure 7, which gives the evolution of the coefficient on lagged inflation in a simple specification regressing inflation as measured by the Consumer Price Index on a constant term, itself lagged, and the unemployment rate, using annual data since 1948, and backward-looking rolling samples of 15 years. It shows the increase in the coefficient early on, the long period of stability around 1, and the sharp recent decline. The coefficient today is not significantly different from zero.

A small detour: Another dimension of change of the Phillips curve is not directly relevant to the issue at hand, but is much discussed and must be mentioned. Since

the mid-1990s, the coefficient measuring the effect of unemployment on inflation has become smaller and less precisely estimated over time (for example, see Blanchard, Cerutti, and Summers 2015; Miles, Panizza, Reis, and Ubide 2017). The origin of this decrease remains largely mysterious. This smaller and imprecise estimate makes it hard to pin down the natural rate of unemployment and raises additional challenges for macroeconomic policy. Some researchers have argued that unemployment no longer has an effect on inflation, at least over some unemployment and inflation range. If it were true, this would have dramatic implications for macroeconomic policy. (For the implications of strict downward rigidity, see Dupraz, Nakamura and Steinsson 2017, and see also Farmer 2013.) I find it difficult to believe that a tighter labor market does not lead to more upward pressure on desired real wages, and in turn, given expected inflation, upward pressure on nominal wage inflation. Indeed, I read the evidence as suggesting that the effect of unemployment on wage determination and in turn on wage inflation, while smaller, remains positive.

Returning to the decrease in the coefficient on past inflation, there can be little doubt that it reflects primarily a change in expectation formation: more specifically, that those setting prices and wages now react less to movements in past inflation. However, as the Lucas critique has made clear, even a zero coefficient on past inflation does not imply that there is an exploitable trade-off between unemployment and inflation. Thus, the question is what hides behind this change in expectations.

I can think of two explanations. First, more stable inflation expectations may arise from increased credibility of monetary policy. Monetary policy may be more credible because of the adoption of inflation targeting, a more explicit target for inflation, and the decrease in the standard deviation of inflation. Second, the experience of low and stable inflation may mean that it is no longer salient, and movements in inflation are ignored by wage- and price-setters. To quote Alan Greenspan (2001): "Price stability is best thought of as an environment in which inflation is so low and stable over time that it does not materially enter into the decisions of households and firms."

Which of these explanations is more relevant has important implications for the natural rate hypothesis. Under the first explanation, any attempt by the central bank to decrease unemployment below the natural rate and, in doing so, increase core inflation, will decrease credibility and lead to an adjustment of expectations. Under the second, the central bank may be able to decrease unemployment and increase inflation without affecting expectations, so long as inflation remains low enough not to become salient.

How can one tell which hypothesis is more relevant? If credibility of the inflation target is the underlying explanation, then inflation expectations should respond more to core inflation, and less to deviations of headline inflation from core. (Headline inflation, which includes food and energy prices, is more volatile.) If instead, decreased salience is the reason, one should find that inflation expectations respond little to core, but respond to deviations of headline from core, coming for example from sharp, and thus more salient, changes in gas prices.

Table 1

Regressions of Inflation Expectations of Professional Forecasters and Consumers on Core and Headline Inflation

	1981Q3 to 1995Q4		1996Q1 to 2016Q1	
	Survey of Professional Forecasters	Michigan Surveys of Consumers	Survey of Professional Forecasters	Michigan Surveys of Consumers
Core	0.498***	0.375***	0.547***	–0.111
	[0.038]	[0.061]	[0.052]	[0.125]
Headline minus core	0.125	0.288**	0.077**	0.231***
	[0.099]	[0.093]	[0.029]	[0.060]
Constant	2.024***	1.873***	1.098**	3.134***
	[0.174]	[0.267]	[0.103]	[0.244]
Observations	58	58	83	83
R^2	0.75	0.66	0.60	0.19

Note: The table shows the results of regressions of inflation expectations on core and headline inflation. It looks at two measures of inflation expectations: the forecast of one-year-ahead inflation as reported by the Survey of Professional Forecasters (columns 1 and 3), and the forecast constructed from the Michigan Surveys of Consumers (columns 2 and 4). The first explanatory variable is a four-quarter moving average of core inflation—the rate of change on the Consumer Price Index excluding energy and food prices. The second explanatory variable is the four-quarter moving average of headline inflation minus core inflation. The first two columns look at the subperiod 1981Q3 to 1995Q4, while the last two columns look at the subperiod 1996Q1 to 2016Q1. Robust standard errors in parentheses.
***, **, and * indicate $p < 0.01$, $p < 0.05$, and $p < 0.1$.

Given this motivation, Table 1 shows the results of regressions of inflation expectations on core and headline inflation. It looks at two measures of inflation expectations: the forecast of one-year-ahead inflation as reported by the Survey of Professional Forecasters (columns 1 and 3), and the forecast constructed from the Michigan Surveys of Consumers (columns 2 and 4).[7] (One wishes that there was a corresponding series of inflation forecasts held by firms, but such a series does not exist.) The first explanatory variable is a four-quarter moving average of core inflation—the rate of change on the Consumer Price Index excluding volatile energy and food prices. The second explanatory variable is the four-quarter moving average of headline inflation minus core inflation. The first two columns look at the subperiod 1981Q3 to 1995Q4, while the last two columns look at the subperiod 1996Q1 to 2016Q1. The basic results in the table are robust to using two-quarter to eight-quarter averages, and to dividing the sample at any date in the 1990s.

[7] The questions asked of consumers are: During the next 12 months, do you think prices in general will go up, or go down, or stay where they are? If people answer "up" or "down," they are then asked, "By about what percent do you expect prices to go (up/down) on the average, during the next 12 months?" If they give an answer greater than 5 percent, they are probed to make sure they understood the question. The details of aggregation are given in Curtin (1996).

The regression results suggest two conclusions. First, professional forecasters put more weight on core than on the deviation of headline from core. In the more recent sample, the weight on core has increased and the weight on the deviation has decreased, suggesting indeed higher credibility of monetary policy. Second, consumers, instead, put more weight on the deviation of headline minus core than on core. In the more recent sample, they appear not to put any weight on core (I have no ready explanation for the negative, but insignificant, coefficient on core), and some weight, although less than before, on the deviation of headline from core. This is suggestive of decreased salience: consumers now ignore inflation unless some large change, such as a change in gas or food prices, takes place.

To summarize: The econometric relation between unemployment and inflation today is at odds with the accelerationist hypothesis, suggesting that inflation expectations have become largely nonresponsive to actual inflation. While increased credibility of policy is clearly a factor, the evidence from consumers' expectations suggests that decreased salience may also be at work. To the extent that these expectations, together with those of firms, are the relevant determinants of wage and price decisions, then, so long as inflation remains low enough, there may be an exploitable persistent, if not permanent, trade-off between unemployment and inflation.

Policy Implications and Conclusions

The policy implications of deviations from the natural rate hypothesis depend very much on the specific channels, the nonlinearities, and the asymmetries that each of these channels implies. Persistence based on loss of morale or skills by workers may have different welfare implications from hysteresis based on insider–outsider considerations.[8] Persistence based on the effects of long-term unemployment is more likely to be asymmetric than persistence based on the effects of activity on R&D and technological progress. It is also more likely to be nonlinear with respect to the depth and the length of recessions. At this point, the empirical evidence is just too crude to give us precise guidance.

Yet the basic implications of deviations from either the independence hypothesis or the accelerationist hypothesis, or both, can be shown simply. Start with the independence hypothesis. Assume that (the log of) potential output, y^* follows

$$y^*(+1) = ay^* + b(y - y^*), \text{ where } a \leq 1$$

Potential output next period, $y^*(+1)$, depends on potential output today and on the deviation of actual output from potential output today. For notational simplicity, the specification ignores all other shocks that affect potential output, and normalizes

[8] Gali (2016) gives a full treatment of policy implications in a model where hysteresis is derived from insider–outsider considerations. See also the analysis of optimal monetary policy in an insider–outsider model by Alogoskoufis (2017).

long-run potential output, if the deviation of output from potential is equal to zero, to be equal to zero.

The parameter b captures the effect of the output gap on potential output, and the parameter a captures the persistence of the effect. Under the strict independence hypothesis, b is equal to zero. Under the strict hysteresis hypothesis (namely that the effect of the output gap on potential output is permanent), b is positive and a is equal to one. I have argued however that these two cases are too extreme. In most models (and in reality), b is likely to be positive and a to be less than one. We can think of the independence hypothesis as small values of b and a, and the hysteresis hypothesis as large values of b and a.

Turn to the accelerationist hypothesis. Assume that the relation between inflation and output is given by:

$$\pi = c(y - y^*) + E\pi$$
$$\text{where } E\pi = 0 \text{ for } -x \leq \pi \leq x, \pi(-1) \text{ otherwise,}$$

where $\pi(-1)$ is the rate of inflation last period. The rate of inflation π depends on the deviation of output from potential, and on expected inflation. Salience is captured by the parameter x. So long as inflation or deflation is smaller than x, expected inflation is constant, normalized here to zero. If inflation or deflation exceed x, inflation or deflation become salient, and is assumed to be equal to lagged inflation. Thus, deviations from the accelerationist hypothesis are captured by positive values of x.

Now consider the trade-off between inflation and output under different assumptions. Suppose first that both the independence hypothesis and the accelerationist hypothesis strictly hold, so b in the first equation and x in the second equation are equal to zero. Consider a one-period increase in output gap, $y - y* \equiv \Delta > 0$. From the second equation, this one-period increase leads to a permanent increase in inflation of $c\Delta$, an unappealing trade-off.

Relax the independence assumption, so b and a are now positive. The one-period increase in the output gap now leads to an increase in potential output in future periods, thus a total increase of $\Delta(1 + b + ab + a^2b + \dots) = \Delta + (b/(1 - a))\Delta$, where the first term reflects the initial output gap and the second reflects the sum of the increases in potential output that result from the initial output gap. The increase in inflation is the same as before, thus equal to $c\Delta$. In short, failure of the independence hypothesis leads to a more appealing trade-off between output and inflation.

Relax instead the accelerationist hypothesis, so x is now positive. Assume past inflation to be equal to zero. As long as the output gap is such that inflation does not exceed $c\Delta$, the increase in the output gap leads to higher current inflation but no increase in inflation in future periods. Thus, failure of the accelerationist hypothesis leads again to a more attractive trade-off between output and inflation.

Relax both hypotheses, and an increase in the output gap today leads to both a larger increase in future output and a smaller increase in future inflation, with both effects leading to an even more attractive trade-off between output and inflation.

This toy model can and should be extended in many dimensions, in particular to allow for a richer specification of the response of inflation expectations to actual inflation, for asymmetric effects of recessions and booms, for the presence of shocks, and for uncertainty about the extent of the deviation from the natural rate hypothesis. The general conclusion is likely to remain the same: Failure of either of the hypotheses leads to a more attractive trade-off between output and inflation, and, in the presence of shocks, suggests a stronger role for stabilization policy. If the independence hypothesis fails, adverse shocks are more costly, and stabilization policy more powerful. If the accelerationist hypothesis fails, there is more room for stabilization policy to be used at little inflation cost.

Where does this leave us? It would be good to have a sense of the values of a, b, c, and x, or more generally, a sense of the specific channels at work. The empirical part of this paper has shown that we are still far from such an understanding. Thus, the general advice must be that central banks should keep the natural rate hypothesis (extended to mean positive but low values of b and a) as their baseline, but keep an open mind and put some weight on the alternatives. For example, given the evidence on labor force participation and on the stickiness of inflation expectations presented earlier, I believe that there is a strong case, although not an overwhelming case, to allow US output to exceed potential for some time, so as to reintegrate some of the workers who left the labor force during the last ten years.

■ *I thank the editors for their suggestions, Larry Summers for many discussions, David Autor, David Cho, Jordi Gali, Egor Gornostay, Alan Krueger, and Mathias Trabandt for data and help, Marios Angeletos, Larry Ball, Olivier Coibion, Nicola Gennaioli, and Robert Solow for comments, and Julien Acalin, Thomas Pellet, and Colombe Ladreit for excellent research assistance. An appendix with methodological details and further results is available with this paper at http://e-jep.org.*

References

Aaronson, Stephanie, Tomaz Cajner, Bruce Fallick, Felix Galbis-Reig, Christopher Smith, and William Wascher. 2014. "Labor Force Participation: Recent Developments and Future Prospects." *Brookings Papers on Economic Activity*, Fall, pp.197–272.

Abraham, Katherine G., John C. Haltiwanger, Kristin Sandusky, and James Spletzer. 2016. "The Consequences of Long Term Unemployment: Evidence from Matched Employer–Employee Data." NBER Working Paper 22665.

Alogoskoufis, George. 2017. "The Clash of Central Bankers with Labor Market Insiders, and the Persistence of Inflation and Unemployment." *Economica*, Early view article, May 15. http://onlinelibrary.wiley.com/doi/10.1111/ecca.12241/full.

Anzaotegui, Diego, Diego Comin, Mark Gertler,

and Joseba Martinez. 2016. "Endogenous Technology Adoption and R&D as Sources of Business Cycle Persistence." NBER Working Paper 22005, February.

Autor, David H., and Mark G. Duggan. 2006. "The Growth in the Social Security Disability Rolls: A Fiscal Crisis Unfolding." *Journal of Economic Perspectives* 20(3): 71–96.

Ball, Laurence. 2009. "Hysteresis in Unemployment." March. *Understanding Inflation and the Implications for Monetary Policy: A Phillips Curve Retrospective*, edited by Jeff Fuhrer, Yolanda K. Kodrzycki, Jane Sneedon Little, and Giovanni P. Olivei. Federal Reserve Bank of Boston and MIT Press.

Ball, Laurence M. 2014. "Long-Term Damage from the Great Recession in OECD Countries." *European Journal of Economics and Economic Policies: Intervention* 11(2): 149–60.

Blanchard, Olivier, Eugenio Cerutti, and Lawrence Summers. 2015. "Inflation and Activity—Two Explorations, and their Monetary Policy Implications." ECB E-Book, *Inflation and Unemployment in Europe, 2015*. ECB Forum on Central Banking, Conference proceedings, 21–23, May, 2015, Sintra, Portugal. www.ecb.europa.eu/pub/pdf/other/ecbforumoncentralbanking2015en.pdf.

Blanchard, Olivier, and Peter Diamond. 1994. "Ranking, Unemployment Duration and Wage Determination." *Review of Economic Studies* 61(3): 417– 34.

Blanchard, Olivier J., and Lawrence H. Summers. 1986. "Hysteresis and the European Unemployment Problem." *NBER Macroeconomics Annual*, vol. 1, edited by Stanley Fischer, 15– 78. MIT Press.

Blanchard, Olivier, and Justin Wolfers. 2000. "The Role of Shocks and Institutions in the Rise of European Unemployment: The Aggregate Evidence." *Economic Journal* 110(462): 1–33.

Caballero, Ricardo J., and Mohamed L. Hammour. 1994. "The Cleansing Effect of Recessions." *American Economic Review* 84(5): 1350–68.

Christiano, Lawrence J., Martin S. Eichenbaum, and Mathias Trabandt. 2016. "Unemployment and Business Cycles." *Econometrica* 84(4): 1523–69.

Cœuré, Benoît. 2017. "Scars or Scratches? Hysteresis in the Euro Area." Speech given at the International Center for Monetary and Banking Studies, Geneva, May 19.

Curtin, Richard. 1996. "Procedure to Estimate Price Expectations." Unpublished paper, University of Michigan, January.

de Grauwe, Paul, and Cláudia Costa Storti. 2004. "The Effects of Monetary Policy: A Meta-Analysis." CESifo Working Paper 1224, June.

Dupraz, Stéphane, Emi Nakamura, and Jón Steinsson. 2017. "A Plucking Model of Business Cycles." Unpublished paper, Columbia University, May.

Farmer, Roger. 2013. "The Natural Rate Hypothesis: An Idea Past Its Sell-by Date." *Bank of England Quarterly Bulletin* 53(3): 244–56.

Fernald, John G. 2015. "Productivity and Potential Output Before, During, and After the Great Recession." Chap. 1 in *NBER Macroeconomics Annual 2014*, Vol. 29.

Foster, Lucia, Cheryl Grim, and John Haltiwanger. 2014. "Reallocation in the Great Recession: Cleansing or Not?" NBER Working Paper 20427, August.

Friedman, Milton. 1968. "The Role of Monetary Policy." Presidential address delivered at the 80th Annual Meeting of the American Economic Association. *American Economic Review* 58(1): 1–17.

Gali, Jordi. 2016. "Insider-Outsider Labor Markets, Hysteresis and Monetary Policy." April. Universitat Pompeu Fabra working Paper 1506.

Greenspan, Alan. 2001. "Transparency in Monetary Policy." October. Remarks at the Federal Reserve Bank of St. Louis, Economic Policy Conference, St. Louis, Missouri (via videoconference), October 11.

Harding, Don, and Adrian Pagan. 2002. "Dissecting the Cycle: A Methodological Investigation." *Journal of Monetary Economics* 49(2): 365–381.

Krueger, Alan B., Judd Cramer, and David Cho. 2014. "Are the Long-Term Unemployed on the Margins of the Labor Market?" *Brookings Papers on Economic Activity*, Spring, pp. 229–299.

Miles, David, Ugo Panizza, Ricardo Reis, and Ángel Ubide. 2017. *And Yet It Moves: Inflation and the Great Recession*. Geneva Reports on the World Economy, no. 19. International Center for Monetary and Banking Studies (ICMB) and Center for Economic Policy Research (CEPR).

Perry, George L. 1970. "Changing Labor Markets and Inflation." *Brookings Papers on Economic Activity* no. 3, pp. 411–48.

Perry, George L. 1978. "Slowing the Wage-Price Spiral: The Macroeconomic View." *Brookings Papers on Economic Activity*, no. 2, pp. 259–99.

Phelps, Edmund S. 1972. *Inflation Policy and Unemployment Theory: The Cost–Benefit Approach to Monetary Planning*. London: MacMillan.

Summers, Lawrence. 2014. "Fiscal Policy and Full Employment." April. Speech at Center for Budget and Policy Priorities Event on Full Employment, April 2.

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

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 121–134

Short-Run and Long-Run Effects of Milton Friedman's Presidential Address

Robert E. Hall and Thomas J. Sargent

The centerpiece of Milton Friedman's (1968) presidential address to the American Economic Association, delivered in Washington, DC, on December 29, 1967, was the striking proposition that monetary policy has no longer-run effects on the real economy. Friedman focused on two real measures, the unemployment rate and the real interest rate, but the message was broader—in the longer run, monetary policy controls only the price level. We call this the monetary-policy invariance hypothesis.

By 1968, macroeconomics had adopted the basic Phillips curve as the favored model of correlations between inflation and unemployment. Unemployment was taken as a good measure of slack or tight conditions. In a slack economy, sellers would gradually cut their prices, and in a tight one, they would gradually raise them. Friedman's presidential address was commonly interpreted as a recommendation to add a previously omitted variable, the rate of inflation anticipated by the public, to the right-hand side of what then became an augmented Phillips curve. Friedman's emphasis on this additional variable was distinctive, but not new. Some years earlier, Samuelson and Solow (1960) had observed that the Phillips curve could shift in ways that depended on a number of factors, including the public's expectations about

■ *Robert E. Hall is Robert and Carole McNeil Joint Hoover Senior Fellow and Professor of Economics, Stanford University, Stanford, California. Thomas J. Sargent is the W. R. Berkley Professor of Economics and Business, New York University, New York City, New York, and Senior Fellow, Hoover Institution, Stanford, California. Their email addresses are rehall@stanford.edu and thomas.sargent@nyu.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.1.121

doi=10.1257/jep.32.1.121

future inflation. Phelps (1967) offered a similar analysis shortly before Friedman, and is often cited in conjunction with Friedman's presidential address.

Friedman's monetary-policy invariance hypothesis implies sharp differences in responses of unemployment to different types of monetary impulses. If a monetary expansion was unanticipated, unemployment would decrease. On the other hand, if a monetary impulse was fully anticipated, there would be no response of unemployment—all of the response will take the form of a change in inflation. A persistently expansionary monetary policy—and therefore a monetary policy expected to be expansionary—would raise anticipated inflation and in this way shift the Phillips curve upward. If the shift was complete, the invariance hypothesis would hold. Friedman's presidential address was an admonition to distinguish sharply between short-run and long-run effects of monetary policy.

We believe that Friedman's main message, the invariance hypothesis about long-term outcomes, has prevailed over the last half-century based on the broad sweep of evidence from many economies over many years. Subsequent research has modified Friedman's ideas about transient effects and has not been kind to the Phillips curve. But we will argue that Friedman's exposition of the invariance hypothesis in terms of a 1960s-style Phillips curve is incidental to his main message. The evidence makes us believe that the invariance hypothesis has stood up well, even though the Phillips curve has not held up as a structural equation in macro models.

We should note at the outset that we recognize small exceptions to the monetary-policy invariance principle. In economies with non-interest-bearing currency, the rate of inflation influences the real cost of holding currency. We believe that these effects are small enough to neglect in this article.

Friedman's Message in 1968

Friedman (1968) set forth two propositions about monetary policy that immediately stirred controversy, but are now close to settled: "(1) It cannot peg interest rates for more than very limited periods; (2) It cannot peg the rate of unemployment for more than very limited periods" (p. 5). These propositions have come to be known as the natural rate hypotheses about the real interest rate and the unemployment rate: The two variables have natural rates. At most, monetary policy induces only transitory deviations of the real rate and the unemployment rate from their natural rates. We regard these natural rate hypotheses as implications of the more general monetary-policy invariance hypothesis.

Friedman explained the natural real interest rate as follows:

> Let the monetary authority keep the nominal market rate for a time below the natural rate by inflation. That in turn will raise the nominal natural rate itself, once anticipations of inflation become widespread, thus requiring still more rapid inflation to hold down the market rate. Similarly, because of the Fisher

> effect, it will require not merely deflation but more and more rapid deflation to hold the market rate above the initial "natural" rate (p. 8).

With respect to unemployment and the labor market, Friedman wrote:

> The "natural rate of unemployment," in other words, is the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is embedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on (p. 8).

In 1968, the idea of a Phillips curve was ascendant: expansionary monetary policy could drive down the unemployment rate, but at the cost of higher inflation. A tradeoff was thought to exist, even in the longer run. Economies willing to accept more inflation could have tighter labor markets with high employment and lower unemployment. Friedman explained that such a tradeoff would not occur in the longer run:

> [T]here is always a temporary trade-off between inflation and unemployment; there is no permanent trade-off. The temporary trade-off comes not from inflation per se, but from unanticipated inflation, which generally means, from a rising rate of inflation. The widespread belief that there is a permanent trade-off is a sophisticated version of the confusion between "high" and "rising" that we all recognize in simpler forms. A rising rate of inflation may reduce unemployment, a high rate will not.

To put it differently, Friedman was arguing that expected inflation was a determinant of actual inflation. We believe that Friedman's decision to express the monetary policy invariance hypothesis in terms of expected inflation on the right-hand side of the Phillips curve was understandable at a time when the Phillips curve had recently acquired a firm grip on thinking about inflation. But we also believe that it pointed the profession toward a limited view of the interactions between inflation and real outcomes.

The Profession's Early Reaction

Adding Expected Inflation to the Phillips Curve

Evaluation of Friedman's formulation that expected inflation shifted the Phillips curve began immediately. Solow (1968, pp. 10–11) and Tobin (1968 pp. 48–54.) added a forecast of inflation to the right-hand sides of their Phillips curves, with a measure of inflation on the left-hand side and unemployment on the right-hand side. Gordon (1970) soon followed. In these papers, the forecast of inflation was

modeled as a distributed lag on past inflation, with lag coefficients that summed to one. Their general finding was that the forecast of inflation received a coefficient of less than one, which led these early investigations to conclude that Friedman was only partly right: they inferred that the Phillips curve shifts upward by only a fraction of expected inflation, so although the long-run Phillips curve is steeper than the short-run curve, it is not vertical. And it is a vertical Phillips curve that expresses the invariance hypothesis, in the interpretation of these authors.

Lucas (1972a) and Sargent (1971) observed that the finding in the distributed-lag-forecast literature was not evidence against the invariance hypothesis. They noted that, under a monetary policy that delivers mean-reverting inflation, the optimal distributed lag forecast will embody the tendency for inflation to subside after a rise. A coefficient of less than one on a distributed lag with coefficients summing to one is the expected outcome in the likely case of mean reversion. The mistake in the distributed-lag approach was to assume that the coefficients in the distributed lag summed to one. That choice amounted to assuming the inflation was a persistent random walk, contrary to the evidence that inflation had been mean-reverting. Because mean-reversion of inflation could be measured in a companion equation, the early studies failed to make full use of the available data.

Further, Lucas (1972a) and Sargent (1971) observed that the problem with the first test of the invariance hypothesis—the failure to take account of the best way to construct a forecast of inflation—was an example of a much more general principle for testing models containing expectations. The principle is rational expectations. Models of expectation formation should not be based on imputing simple-minded ways that people form expectations, such as extrapolating past observations. Rather, econometricians should apply the same standards of rational behavior to the formation of expectations that they do to other aspects of economic choices. Lucas and Sargent recommended tests of Friedman's invariance hypothesis based on rational expectations instead of the model of expectations assumed by Solow, Tobin, and Gordon. But the rational expectations assumption was foreign to macroeconometric practice in that era.

Notice that the critique had two layers: 1) people may forecast inflation by applying lag coefficients to past inflation, but if they do, it would not be rational to use coefficients that sum to one if inflation was less persistent than a random walk; and 2) it's unlikely that expectations would consider only lagged values of inflation—for example, if an inflation hawk has just taken over the central bank, people might reasonably expect a larger decline in inflation than indicated by a previously successful distributed lag equation. Sargent had a colorful way to drive the second point home. Suppose, he asked, that the rules of American football were changed so that the offense had the ball for only three downs rather than four. Prior to the change, nobody would expect a team to punt on third down. After the change, the rational fan would expect frequent punts on third down. Historical punting tactics would not be a rational guide to tactics under the new rule.

Rational Expectations

Friedman's (1968) presidential address, along with Phelps (1967), drew the attention of young researchers to an important part of macroeconomic theory that was unfinished in 1968—how to build a model of expectations formation that was consistent both with optimizing behavior and the structure of a macroeconomic model. In response, they rolled up their sleeves and learned the mathematics and probability theory required to apply the rational expectations hypothesis of Muth (1961) in macroeconomic models. Before Phelps and Friedman, the rational expectations hypothesis, if considered at all, was just one of several possible assumptions about expectations that an econometrician could use. The most popular model asserted that expectations were adaptive—people extrapolated recent behavior of a variable in a fixed way to form an expectation of its future values. The mathematics of prediction theory used by Muth (1961), and the idea of fixed points in function spaces underlying Muth's analysis, were unfamiliar to most macroeconomists. That changed soon after Friedman's presidential address.

Lucas (1972b) used the rational expectations hypothesis to produce a striking clarification and strengthening of Friedman's invariance hypothesis. Lucas's paper offered one of the first rigorous developments of a general equilibrium model that imposed Muth's rational expectations assumption. Lucas's notion of rational expectations, and a huge successor literature, starts by conceiving of a model as a joint probability distribution over sequences of exogenous processes and choices. It then posits that the agents in the model also use the model itself to make inferences about the future behavior of variables relevant to their decisions. In a "communism of models" comprising 1) the agents in the model, 2) nature, and 3) the model builder—all three share the same statistical model. This simplifying assumption sharpens and focuses the analysis. In Lucas's model, agents are imperfectly informed about random changes in the money stock. That causes agents to be only imperfectly able to distinguish outcomes caused by money supply changes, on the one hand, and the real determinants of employment and output, on the other hand. They make decisions that are optimal given their information limitations but recurrently mistaken relative to those that would be made with full information. The limitations on information cause monetary changes to affect real variables. Real variables in this framework do not respond to the systematic, predictable component of the money supply. Thus, Lucas produced a formal, rigorous expression of Friedman's invariance principle. Two otherwise similar economies having the same money shocks but differing with respect to the predictable parts of money growth will have the same output and employment movements, and will differ only in their rates of inflation.

In Lucas's (1972b) and other general equilibrium models of money, it matters how a government induces changes in the money supply. Most of the ways that a government injects or withdraws money are partly fiscal policies and are not neutral—they affect output and other real variables through fiscal channels—even if they are foreseen. To create an explicit framework in which foreseen monetary shocks are neutral, the government in Lucas's model hands out money in a very

special way: namely, proportionally to agents' initial holdings of money each period. These transfers are, accordingly, equivalent to a pure change in the units in the monetary standard. To disentangle real from monetary shocks, the agents in Lucas's model solve a signal extraction problem. Agents know joint probability distributions and use Bayes's law to solve the signal extraction problem arising from their limited information. In this way, Lucas transformed Friedman's informal distinction between the long run and the short run into a tight mathematical distinction between predictable and unpredictable policies and outcomes.

Lucas did not ask how agents inside his model might have learned about a rational expectations equilibrium. They just do—they are born knowing the relevant probability distributions. They do not need Bayes's Law to improve their knowledge of the model. Perhaps their ancestors successfully resolved model uncertainty by applying Bayes's Law. Researchers in the 1980s took up the question of whether agents who don't know the model might learn about it by applying an adaptive algorithm or some version of Bayes's Law in settings with model uncertainty. That literature described convergence theorems in the form of conditions under which a self-referential system comprised of agents who initially do not know enough to do what they are supposed to do inside a rational expectations equilibrium could converge to a rational expectations equilibrium. Sargent (1999) summarized this literature and described how it applies to the issues raised in Friedman's presidential address, the analysis of Lucas, and Kydland and Prescott (1977) and other contributors to this branch of macro theory. The literature on learning about a rational expectations equilibrium relies heavily on a theory of stochastic approximation that uses simulations to maximize an unknown function. In fact, Friedman and Savage (1947) was an early technical contribution to that literature.

Although there are now serious applications of the literature on learning to macro policy formulation, it nevertheless remains the case that most policy models today are formulated under the communistic rational expectations principle that all agents use the author's model in solving their optimization and forecasting problems.

Lucas (1973) carried out an empirical investigation in the rational expectations framework, with emphasis on the invariance hypothesis. He studied panel data on inflation and unemployment across countries and years. His concept of invariance was more general than just comparing policies of high and low inflation—in the long run, real outcomes such as unemployment are invariant to all types of differences in monetary policy. He summarized the framework this way: "These data are examined from the point of view of the hypothesis that average real output levels are invariant under changes in the time pattern of the rate of inflation, or that there exists a 'natural rate' of real output." His findings gave strong support to the invariance hypothesis. In particular, high-inflation countries did not have lower unemployment.

The NAIRU and the Acceptance of the Natural Rate Hypothesis

Economists who initially questioned Friedman's monetary policy invariance hypothesis, notably Modigliani and Papademos (1975), came around to, at least,

a more-limited version of it within a decade. One implication of the hypothesis is that, at the natural rate of unemployment, if inflation is replicating itself, and the price level is neither accelerating nor decelerating, the unemployment rate will be at its natural level. On this basis, some of the former skeptics renamed the natural rate the "non-accelerating inflation rate of unemployment" or NAIRU. This brand distinction followed a tribal distinction between "saltwater" and "freshwater" macroeconomics described in Hall (1976). It is unfortunate that many commentators have misconstrued Hall's tongue-in-cheek account of schools of macroeconomics as indicating a broader schism between coastal and mid-west approaches to macroeconomics. No such schism existed or exists among researchers actually working in the research trenches. Macroeconomists have their disagreements, of course, but they share beliefs about equilibrium concepts, analytical tools, and salient observations, and all have gathered insights and inspirations from great predecessors such as Frank Ramsey, John Hicks, Kenneth Arrow, Milton Friedman, and John Maynard Keynes.

A custom related to the term NAIRU was to use the term "accelerationist" to describe a related hypothesis that Friedman considered—that an attempt to hold unemployment below the natural rate with monetary policy would result in ever-accelerating inflation. The corollary, that a monetary policy that generated ever-higher inflation would keep unemployment below the natural rate, is a violation of the monetary-policy invariance hypothesis. We are not aware that any believer in the NAIRU has advocated such a policy, however.

In recent decades, the idea of a natural rate or NAIRU has become uncontroversial. Controversy has shifted to debates over the level of the natural rate and how to model the inflationary process in other respects.

Commitment Issues in Monetary Policy

The arrival of rational expectations in economics focused attention on the importance of timing protocols in the analysis and design of macroeconomic policies. With forward-looking agents who anticipate future policy decisions, equilibrium outcomes depend sensitively on who knows and chooses what when. A natural consequence was to define economic policies more tightly as decision rules stating planned responses to possible future events. Analytical tools of backward induction and dynamic programming came to macroeconomics. Notions of short run and long run were sharpened, and economists came to understand the role of consistency over time. Although Milton Friedman had earlier played an important role in the invention of sequential analysis and dynamic programming—see Friedman and Friedman (1998, pp. 137–39) and the introduction to Wald (1947)—he did not use them in his macroeconomic research.

Kydland and Prescott (1977) and Barro and Gordon (1983) analyzed the consequences of alternative timing protocols for monetary policy. They compared outcomes in economies where the central bank is free to make policy on the spot, unable to commit to a policy in advance, with ones in which a time-zero central bank could choose once and for all. They took the "on the spot" timing protocol to

be the one in place in practice. If there is an advantage to creating a positive inflation surprise, the central bank faces a temptation to inflate more than expected. Kydland and Prescott concluded that the central bank would give in to that temptation. In this case, the rational-expectations equilibrium involves inflation rates high enough to prevent the central bank from creating even more inflation as a surprise. Barro and Gordon applied a theory of reputation to describe a better (subgame perfect) equilibrium where fear of losing its reputation for noninflationary policy blocks the perverse equilibrium.

Maybe it was a coincidence, but by about 1990, central banks around the world almost universally stopped inflationary policies. In the last quarter-century or so, high rates of inflation have arisen only in extraordinary circumstances, like the period of the transition economies that arose in the aftermath of the breakup of the Soviet Union, or in cases of comprehensively failed states like Zimbabwe and more recently Venezuela. Other countries now having high inflation rates are poorly governed and rely heavily on central bank borrowing to finance their governments.

Later Responses to the Presidential Address: The Search for a Theory of the Phillips Curve

Friedman (1968) convinced multiple generations of macroeconomists that the two forces driving inflation were market tightness and expected inflation. The expectation-augmented Phillips curve became a standard feature of the general equilibrium models used by central banks and other policymakers. As the macro profession focused more on formal modeling with microeconomic foundations, a search began for a specification of the Phillips curve that appeared to satisfy these advancing standards.

The general equilibrium model resulting from this process took the general form of a three-equation model, comprising the Phillips curve; an IS curve relating output negatively to the real interest rate; and a Taylor rule, describing how the central bank provides a nominal anchor by setting the interest rate to achieve a target inflation rate in the longer run. Woodford (2003) is a canon of this literature. In that model, the Phillips curve is an equation with inflation as the left-hand variable and two right-hand variables: 1) unemployment or another slack-versus-tightness measure; and 2) the mathematical expectation of future inflation, derived from the model itself. Most research in this framework adheres to the principle of communistic rational expectations.

Calvo (1983) was a key step in the process of formalizing modern Phillips-curve theories based on explicit models of sticky prices. That paper led to what came to be called the New Keynesian family of general equilibrium macro models. Calvo hypothesized that sellers kept their prices fixed until a random event occurred that freed them from the stale prices and allowed them to set a new price. Sellers needed to form expectations of conditions in their markets over the indefinite future to

figure out how to set prices that would remain in place in the future. Although the model can be written out in an extended form in which sellers have expectations about the future demand functions that will determine future sales (and thus output), the custom from the start has been to restate the model in the form suggested by Friedman (1968), where expected future prices stand in for the future demand functions. The logic is that future prices will be set, in part, by firms that have just been freed from their sticky prices.

The Calvo (1983) setup differs fundamentally from the idea popular in 1968 that expected inflation was a distributed lag on past inflation. Sellers in the Calvo model are forward-looking. The model is capable of addressing questions about changes in monetary policy regimes, where the backward-looking model stumbles for reasons explained in Lucas (1976) and captured in Sargent's football analogy. A change in monetary policy changes the coefficients of a forecast based on a distributed lag of past inflation.

Variants of the Calvo (1983) model dominate practical macro models today. Their common idea is that sellers put their prices on autopilot between occurrences that arise at random times and cause sellers to think through pricing more fully. A basic asymmetry runs through this line of work. The autopilot governs prices between these occurrences. Buyers have a call option, in effect, on the seller's output. One could instead imagine that a seller puts output on autopilot and lets the market set the price between full resets of output. The New Keynesian paradigm requires this asymmetry by taking it as given that a significant part of the volatility of output reflects product demand fluctuations. With short-run sticky prices, the call-option setup implies that movements in output are bigger than they would be with flexible prices. The flexibility of prices absorbs demand changes and thus reduces the response of output to the demand changes.

What we are referring to as the call-option property of New Keynesian models is also responsible for the role of unemployment or other tightness/slack measures in the Phillips curve. The initial effect of a decline in demand is a slacker market, with lower output and higher unemployment in the corresponding labor market. If the drop in demand is expected to persist, lower output and higher unemployment will cause sellers to set lower prices in the future, so market slackness predicts lower inflation. In this way, the autopilot that keeps a firm's price constant into the future rationalizes the Phillips curve. If the autopilot were instead to stabilize the output of a firm, the firm's price would respond quickly and output would be sticky. The Phillips curve would look completely different.

Our commentary concentrates on the Phillips curve, but we should mention that Friedman's (1968) presidential address assumed that the central bank uses the money supply as an intermediate target. Central banking practice shifted two decades later to using the interest rate as the intermediate instrument of its operating policy. Macroeconomists continue to speak of "monetary policy" and "monetary theory," although money has been pushed into the background in models in the Woodford style. What serves as a nominal anchor in these models is not the purposefully controlled supply of money advocated by Friedman, but rather

a purposeful feedback rule from prices to the real interest rate in conjunction with assumptions that make the price level sticky in particular ways.

The Missing Empirical Relationship between Slackness and Inflation

The Phillips curve originated as an observation of an empirical relationship in UK data, a relationship which seemed to persist in US data in the 1960s. Friedman's (1968) presidential address adopted the assumption that measures of economic slack are inversely correlated with inflation. But under closer examination and with more recent data, this relationship seems weak or nonexistent.

Stock and Watson (2010) take a close look at evidence in US data, including the deep recession years immediately following the financial crisis of 2008. They find no support for the standard Phillips curve property that the rate of change of prices depends on the level of unemployment. Rather, in response to an adverse shock that causes a quick increase in unemployment, which then gradually subsides, the inflation rate falls a bit immediately and then remains constant. If anything, the rate of change in prices depends on the rate of change of unemployment, a relation inconsistent with the Calvo model. In the depression of the US economy starting in 1929, prices and wages fell during the contraction but stopped falling when the contraction ended and the economy appeared to be stagnant. In this symposium, Blanchard also discusses the weakness of the evidence for a slackness effect in the Phillips curve. A study of episodes of major changes in inflation rates is also instructive about the failures of mechanical models of the Phillips curve. Sargent (1982) considers four historical examples in which changes in monetary and fiscal policy regimes resulted in stabilizations following extreme rates of inflation. These reductions in inflation occurred without major slack.

We conclude that the Phillips curve has little value as a component of a model of inflation. It is not a description of the actual behavior of inflation, and it is incapable of dealing with the important question of what happens when macroeconomic policy undergoes major reform. Although Friedman tied the ideas in his presidential address to the Phillips curve, the ideas apply much more generally. In particular, they are central to the analysis of policy regime changes.

Alternative Macro Models for Testing the Invariance Hypothesis

The missing connection between economic slack and inflation represents a challenge for economic analysis. In his own empirical work, Friedman revealed his mistrust of models of short-run dynamics then available, like the simultaneous equations method often associated with the Cowles Commission. Friedman (1970) expressed sympathy with the view that, in the very short run, an assumption of fixed prices may be reasonable, and said that in that case generally accepted Keynesian principles govern the economy. Friedman said that the

challenge was to understand the dynamic transition from the short to the longer run, which in 1970, in his opinion, was not well developed. About this process, Friedman wrote:

> . . . the rate of adjustment in a variable is a function of the discrepancy between the measured and the anticipated [longer-run] value of that variable or its rate of change, as well as, perhaps, of other variables or their rates of change. Finally, I shall let at least some anticipated variables be determined by a feedback process from past observed values (p. 223).

These musings are both insightful and insufficiently precise to guide a tight econometric specification. Subsequent research seeking to use modern methods—like structural vector autoregressions, rational expectations, and recursive formulations of equilibria as tightly parameterized stochastic processes—can be read as showing why Friedman was wise to be cautious.

When confronted with the challenge of doing macroeconomic modeling when causal connections are not clear, it is natural to turn to vector autoregressions that make only limited assumptions about the underlying structure. One enduring influential aspect of Friedman's informal characterizations of the short-run effects of monetary expansions is that, for a while, they drive interest rates and unemployment down. During this period, which lasts several years in many models, inflation rises only slowly. Uhlig (2005) formalized intuitions along the lines of Friedman's in terms of sign restrictions on the coefficients of structural vector autoregressions that would imply this behavior of the model's response to a monetary shock. This approach to measuring responses to shocks continues to play an important role in building structural macroeconometric models—see Christiano, Eichenbaum, and Evans (2005)—and in research on price stickiness.

Another line of research builds dynamic models of price stickiness from data on the prices of individual products. The availability of micro-level data has ignited an active and challenging research program that aims to refine models of price stickiness with an eye to match both the vector autoregression evidence on macroeconomic aggregates and also panel evidence about the price-setting behavior of firms. These models have yielded a wide range of answers. Some agree with the general conclusion of vector autoregressions that the period over which monetary shocks affect real variables is several years. Others, such as Golosov and Lucas (2007), find quantitatively small effects of unanticipated monetary expansions.

We draw two conclusions from this ambitious literature. One is that the features of models needed to replicate the findings based on macroeconomic aggregates, as studied in the vector autoregression literature, are highly specific and therefore fragile—that is, small and seemingly unimportant changes in such models affect the results. Our other lesson is that Friedman's monetary-policy invariance insight is highly robust. Research has not found evidence that monetary policy has a lasting effect on unemployment. The puzzle remains that it is difficult to demonstrate that monetary policy affects inflation either.

The Short- and Long-Run Effects of the Presidential Address

Friedman's (1968) presidential address was aimed at economists, and its effects on world economies operated through the economics profession. Within the profession, the short-run effect was to stimulate many to check Friedman's assertion that, not only did expected inflation matter for actual inflation, it mattered point-for-point in the determination of actual inflation. Within the then-existing framework of the Phillips curve, as Friedman pointed out, the long-run Phillips curve became vertical and the unemployment rate or other measure of slack was invariant to the central bank's inflation choice. The first round of regressions trying to check the idea seemed to show that it was wrong—expected inflation received a coefficient less than one in the early regressions. In the longer run, Friedman's hypothesis of a point-for-point shift of the Phillips curve gained full acceptance among economists.

The more general assertion that real outcomes such as unemployment, employment, and output were invariant to the monetary regime began to be accepted. That idea generalized and replaced the concept of monetary neutrality. Initially, monetary neutrality was thought to apply to the level of the money stock—changes in the stock would ultimately change the price level in proportion and leave real variables unchanged. Some economists interpreted Friedman's (1968) idea as involving neutrality with respect to the money growth rate and the level of inflation. The alternative label for the natural unemployment rate—the non-accelerating rate of inflation unemployment rate (NAIRU)—seems to leave open the possibility of third-degree non-neutrality. In that case, unemployment would be invariant to the price level and the inflation rate, but a central bank could lower unemployment for as long as it wanted by generating a constantly rising rate of inflation. Friedman seemed to leave this issue open, in the passage quoted at the beginning of this article, ending "A rising rate of inflation may reduce unemployment, a high rate will not." With the advent of formal modeling of the issue with rational expectations, first in Lucas (1972b) and later in New Keynesian models, the profession came around to the fuller proposition of the invariance of real outcomes to monetary policies, not just to the level, rate of change, or acceleration of the price level.

The deeper message of Friedman's (1968) presidential address is its extension of the logic of the invariance principle beyond what Friedman described as the long run and in particular to recognize that it is a mistake for policymakers to regard the expected rate of inflation as a determinant of, or anchor for, actual inflation. Rather, the message is that in a coherent model, expected inflation is itself an outcome and that the same fundamentals determine both inflation and the public's expectation of it. Distinguishing the long run from the short run is a handy way to communicate an intuitive version of ideas about the effects of policy, but the advances Friedman stimulated replaced the distinction with a fuller analysis based on optimizing behavior and rational expectations. Macroeconomists today trace the effects of a policy change over time by calculating a function that shows how the response evolves over time following a policy innovation.

Although Muth (1961) had defined and discussed rational expectations almost a decade before the presidential address, the hypothesis had not permeated macroeconomic thinking until the sharp debates unfolded immediately after Friedman's presidential address.

One can trace an intellectual response function to Friedman's innovation: in the first few years, the debate focused on whether Friedman was right that the Phillips curve shifted point for point with expected inflation. That phase ended with the acceptance of that proposition implied by adoption of the NAIRU label by most of Friedman's earlier critics. Then, in the 1970s, the validity of the rational expectations hypothesis was the subject of raging debate. By the 1980s, the hypothesis was mostly accepted, at least as the default way to think about expectations. Authors earn no points for embodying rational expectations in a model, any more than they would for assuming profit- or utility-maximization. Rational expectations is part of the basic conventional toolkit of macroeconomics.

Central banks are responsible for monetary policy in almost all countries. The effects of Friedman's (1968) presidential address on macroeconomic outcomes operated mainly through central bankers. In 1968, and at least through the 1970s, central banking was in a state of deep intellectual confusion. Many central banks behaved as if they lacked tools for managing the rate of inflation. Rather than steering inflation by committing to a monetary rule, as Friedman had recommended well before his presidential address, central banks permitted rising inflation, then endorsed and participated in nonmonetary and harmful policies to try to bring inflation under control. The Federal Reserve, for example, endorsed price controls from 1971 to 1974 and enforced credit controls in 1980.

We believe that Friedman's thinking, expressed in his 1968 presidential address, began a highly successful educational process that led most of the central banks of the world to abandon high-inflation policies and commit to successful inflation-stabilization policies that provided effective nominal anchors. The address itself effectively attacked the idea that low unemployment was a benefit of tolerating high inflation. By calling attention to the roles of forward-looking economic agents, the presidential address laid the foundations for central bankers to believe that commitment to low-inflation policies was key to achieving low inflation. Macroeconomists under Friedman's influence showed central bankers the danger of failing to commit, and the challenge to make commitments credible. The unfavorable experiences in the 1970s around the world resulted from failure to commit, and the successful adoption of more committed policies starting in the 1980s owes a lot to Friedman, much of it channeled through the presidential address.

References

Barro, Robert J., and David B. Gordon. 1983. "Rules, Discretion and Reputation in a Model of Monetary Policy." *Journal of Monetary Economics* 12(1): 101–21.

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

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.

Friedman, Milton. 1968. "The Role of Monetary Policy." Presidential address delivered at the 80th Annual Meeting of the American Economic Association. *American Economic Review* 58(1): 1–15.

Friedman, Milton. 1970. "A Theoretical Framework for Monetary Analysis." *Journal of Political Economy* 78(2): 193–238.

Friedman, Milton, and Rose D. Friedman. 1998. *Two Lucky People: Memoirs.* University of Chicago Press.

Friedman, Milton, and L. J. Savage. 1947. "Planning Experiments Seeking Maxima." In *Selected Techniques of Statistical Analysis for Scientific and Industrial Research, and Production and Management Engineering,* edited by Churchill Eisenhart, Millard W. Hastay, and W. Allen Wallis, 363–372. New York and London: McGraw-Hill.

Golosov, Mikhail, and Robert E. Lucas Jr. 2007. "Menu Costs and Phillips Curves." *Journal of Political Economy* 115(2): 171–99.

Gordon, Robert J. 1970. "The Recent Acceleration of Inflation and Its Lessons for the Future." *Brookings Papers on Economic Activity* no. 1, pp. 8–47.

Hall, Robert E. 1976. "Notes on the Current State of Empirical Macroeconomics." June. Available at https://web.stanford.edu/~rehall/All_publications.htm.

Kydland, Finn E., and Edward C. Prescott. 1977. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy* 85(3): 473–92.

Lucas, Robert E., Jr. 1972a. "Econometric Testing of the Natural Rate Hypothesis." In *The Econometrics of Price Determination: Conference, October 30–31, 1970,* edited by Otto Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.

Lucas, Robert E., Jr. 1972b. "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4(2): 103–24.

Lucas, Robert E., Jr. 1773. "Some International Evidence on Output–Inflation Tradeoffs." *American Economic Review* 63(3): 326–34.

Lucas, Robert E., Jr. 1976. "Econometric Policy Evaluation: A Critique." In *Carnegie-Rochester Conference Series on Public Policy*, vol. 1, pp. 19–46. Elsevier.

Modigliani, Franco, and Lucas Papademos. 1975. "Targets for Monetary Policy in the Coming Year." *Brookings Papers on Economic Activity,* no. 1, pp. 141–65. Spring.

Muth, John F. 1961. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29(3): 315–35.

Phelps, Edmund S. 1967. "Phillips Curves, Expectations of Inflation and Optimal Unemployment over Time." *Economica* 34(135): 254–81.

Samuelson, Paul A., and Robert M. Solow. 1960. "Analytical Aspects of Anti-Inflation Policy," *American Economic Review* 50(2): 177–94.

Sargent, Thomas J. 1971. "A Note on the 'Accelerationist' Controversy." *Journal of Money, Credit and Banking* 3(3): 721–25.

Sargent, Thomas J. 1982. "The Ends of Four Big Inflations." Chap. 2 in *Inflation: Causes and Effects*, edited by Robert E. Hall. University of Chicago Press for the National Bureau of Economic Research.

Sargent, Thomas J. 1999. *The Conquest of American Inflation.* Princeton University Press.

Solow, Robert M. 1968. "Recent Controversies in the Theory of Inflation." In *Proceedings of a Symposium on Inflation*, edited by Stephen Rousseaus. New York University.

Stock, James H., and Mark W. Watson. 2010. "Modeling Inflation after the Crisis." *Proceedings of the Economic Policy Symposium,* Jackson Hole, Federal Reserve Bank of Kansas City, pp. 173–220. Federal Reserve Bank of Kansas City.

Tobin, James. 1968. "Discussion." In *Proceedings of a Symposium on Inflation,* edited by Stephen Rousseaus, p. 48–54. New York University.

Uhlig, Harald. 2005. "What are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure." *Journal of Monetary Economics* 52(2): 381–419.

Wald, Abraham. 1947. *Sequential Analysis.* New York: John Wiley.

Woodford, Michael. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy.* Princeton University Press.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 135–154

Exchange-Traded Funds 101 for Economists

Martin Lettau and Ananth Madhavan

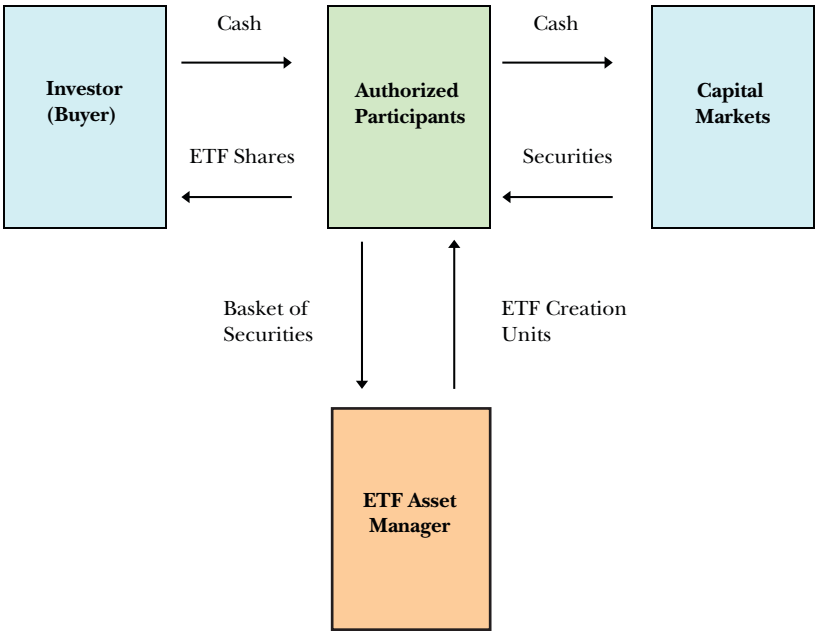
Exchange-traded funds (ETFs) represent one of the most important financial innovations in decades. As such, they are of considerable interest to economists, but the literature on ETFs is, as we shall see, still at an early stage. An ETF is an investment vehicle, with an architecture shown in Figure 1 (to be discussed), that typically seeks to track the performance of a specific index, like an index mutual fund does. But an ETF differs from a mutual fund in fundamental ways, as we will describe below. The first US-listed ETF, the SPDR, was launched by State Street in January 1993 and seeks to track the S&P 500 index. It is still today the largest ETF by far with assets of $178 billion as of September 2017. Following the introduction of the SPDR, new ETFs were launched tracking broad domestic and international indices, and more specialized sector, region, or country indexes. In recent years, ETFs have grown substantially in assets, diversity, and market significance, including substantial increases in assets in bond ETFs and so-called "smart beta" funds that track certain investment strategies often used by actively traded mutual funds and hedge funds. These trends have the potential for dramatically reshaping the broader investment landscape, as we discuss below. Globally, assets

■ *Martin Lettau is the Kruttschnitt Family Chair in Financial Institutions, Haas School of Business, University of California at Berkeley, Berkeley, California. He is a Research Associate at the National Bureau of Economic Research, Cambridge, Massachusetts and a Research Fellow at the Centre for Economic Policy Research, London, United Kingdom. Ananth Madhavan is Managing Director, Global Head of ETF and Index Investing Research, BlackRock, Inc., San Francisco, California. Their email addresses are lettau@berkeley.edu and ananth.madhavan@berkeley.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.1.135

doi=10.1257/jep.32.1.135

Figure 1
The ETF Architecture



of exchange-traded funds under management are $4.3 trillion in September 2017 (exceeding the hedge fund industry) in roughly 6,300 investment vehicles (according to the BlackRock 2017b). These totals should be viewed against the global total market value of equity and fixed income securities in excess of $160 trillion.

In this paper, we begin by describing the structure and organization of exchange-traded funds. We offer a number of contrasts with mutual funds, which are close relatives of exchange-traded funds, describing the differences in how ETFs operate and their potential advantages in terms of liquidity, lower expenses, tax efficiency, and transparency.

We then turn to concerns over whether the rise in ETFs may raise unexpected risks for investors or greater instability in financial markets. Some of the potential issues include what happens when an ETF is delisted; risks when ETFs lend their securities to short-sellers; concerns about ETFs that trade intraday but are based on infrequently traded assets; and whether ETF flows could lead to price distortions or additional volatility. While concerns over financial fragility are worth serious consideration, some of these concerns are overstatements, and for others, a number of rules and practices are already in place that offer a substantial margin of safety.

The conclusion of the article offers some suggestions for future research in this growing field. For more comprehensive treatments of ETFs and related investment vehicles, interested readers might start with Hill, Nadig, and Hougan (2015) and Madhavan (2016).

Structure and Ecosystem: Comparing Exchange-Traded Funds and Mutual Funds

Most economists are familiar with mutual funds, so it is useful to describe how ETFs function by comparing them with mutual funds (for a survey of the literature on mutual funds, see Elton and Gruber 2013).[1]

Who Creates and Trades Shares?

A mutual fund holds the underlying assets: for example, an S&P 500 index fund holds a portfolio of stocks that makes up the S&P 500 index. The manager of a mutual fund will contract with a pricing provider to determine a "net asset value" (NAV) of the fund based on the last recorded prices of the component securities.[2] In a mutual fund, all transactions occur at the end of the day and at net asset value. If this index fund experiences a net in-flow of investment at the end of the day, the mutual fund itself will purchase more shares of stock.

An exchange-traded fund also holds a portfolio of assets; however, in contrast to a mutual fund, it does not interact with capital markets directly. Instead, the ETF manager (or sponsor such as Vanguard or State Street) enters into a legal contract with one or several "Authorized Participants" (APs), typically large financial institutions or more specialized market-makers, who in turn interact with the markets (see Figure 1). In particular, the ETF manager can issue or redeem shares with Authorized Participants in large blocks, known as creation units, in exchange for a basket of securities and/or cash. This mechanism, by which the shares of the ETF are adjusted in response to supply and demand, is known as the creation/redemption mechanism. Here, "creations" refer to increasing the supply of ETF shares; "redemptions" refer to a decrease in the shares outstanding of the ETF.

Both current fund holdings and the basket of securities that the ETF will accept for creations or redemptions on the next business day are published at the end of each trading day. The transactions between an ETF manager and an Authorized Participant are typically either for cash or "in-kind" where the Authorized Participant delivers or receives a basket of securities identical (or very similar) to the ETF's

[1]In particular, we focus here on "open-end" mutual funds, in which the number of "shares" in the fund, and hence its size, can expand and contract. In a closed-end mutual fund, by contrast, the fund's shares, once issued, are fixed in supply and they trade on the open market at prices that could be quite different from net asset value. There is typically a discount on closed-end funds, which represents a puzzle for economists because, in theory, a substantial discount should mean that the fund could benefit its investors by liquidating and returning the proceeds. There is a large literature on the nature and properties of the closed-end fund discount (in this journal, see Lee, Shleifer, and Thaler 1990; see also Lee, Shleifer, and Thaler 1991; Elton, Gruber, and Busse 1998; Berk and Stanton 2007).

[2]For international mutual funds, net asset value is often adjusted, or "fair valued," based on market movements in other markets (for example, by adjusting valuations in emerging markets based on US futures market movements) to prevent gaming. Similarly, bond fund valuations may also be adjusted by the pricing provider because component securities might have traded days, even weeks, ago. Grégoire (2013) finds evidence that mutual funds do not fully adjust their valuations to reflect fair value, and returns remain predictable.

holdings. Like other investors, Authorized Participants can buy or sell ETF shares in the secondary market exchange, but they also can purchase or redeem shares directly from the ETF if they believe there is a profit opportunity. The process of ETF share creation or redemption for an ETF is illustrated in Figure 1, where we show the "in-kind" exchange of securities for ETF shares. The process of a cash creation (not typical) is similar.

Early ETFs were almost exclusively seeking to track broad value-weighted equity indices (for example, the S&P 500) but ETFs today track a wide variety of equity and fixed-income indices. There are also active ETFs that are akin to active mutual funds in that they seek to outperform a benchmark index, but to date they are still a small fraction of total ETF assets.

How is the Price of Shares Determined?

Managers of exchange-traded funds are, like mutual funds, required by the Securities and Exchange Commission (SEC) to publish a "net asset value" for their funds. In contrast to mutual funds, investors in exchange-traded funds mostly do not trade the fund directly. Instead, they deal with each other on an exchange, or with Authorized Participants and other liquidity providers. Investors can buy and sell shares in ETFs through a broker, just as they buy and sell shares of publicly listed companies. This secondary market trading does not lead to transactions in the underlying securities, which greatly reduces the transaction costs that arise when investors redeem from the fund. The secondary market (exchange-traded) trading volume for most ETFs is typically a multiple of the volume of creation/redemption activity by the Authorized Participants. According to Investment Company Institute statistics for 2014, this ratio is about 4:1 over all ETFs.

Although shares of exchange-traded funds can be created or redeemed at the end of each trading day, the Authorized Participants will typically lock in any profits intraday. For example, when an ETF is trading at a premium to an Authorized Participant's estimate of value (which need not be the net asset value of the fund), the Authorized Participant may choose to deliver the creation basket of securities in exchange for ETF shares, which in turn it could elect to sell or keep. The creation/redemption mechanism works through arbitrage to help keep the price of an exchange-traded fund close to the intrinsic value of an ETF's holdings in the underlying market.

In the context of an exchange-traded fund, deviations of price from the announced net asset value do not necessarily imply the existence of arbitrage opportunities, especially for international funds and for funds whose constituents may be difficult to value because of infrequent trading. As noted above, the ETF sponsor contracts with market data vendors (or other third parties) to calculate and publish net asset value based on past prices. Vendors also provide an Intraday Indicative Value that is disseminated at regular intervals during the trading day, typically every 15 seconds. This value is usually based on the most recent (possibly stale) trade. Thus, if the exchange-traded fund holds Japanese stocks, say, the closing price (or quote) from Tokyo is used throughout the US trading day and a foreign exchange

adjustment is made for any change in the yen/dollar relationship since the Tokyo markets are closed. For fixed-income funds, the provider of the Intraday Indicative Value may not necessarily fully update the prices of securities that do not trade, or include adjustments for accrued fees or liabilities that vendors usually reflect in their end-of-day net asset value.

Madhavan and Sobczyk (2016) develop and test a model of exchange-traded fund price dynamics where arbitrage corrects deviations between the price of ETFs and the underlying value of the basket. In their model, the actions of arbitragers reduce these deviations over time, yielding a metric for the speed of price discovery. The model explains why premiums and discounts to net asset value need not necessarily constitute mispricing or the existence of arbitrage opportunities, as well as why ETF returns may be more volatile than the returns of the benchmark index. They empirically estimate the model for the universe of US-listed exchange-traded funds and find that, on average, the speed of price discovery (measured by the half-life to correct any given deviation of price from basket value) is shortest for US equity-focused funds and greatest for international-bond funds, which is consistent with the observed pattern of liquidity.

Ultimately, the intraday tradability of exchange-traded funds is really a by-product of having the price of the fund determined by the market through the interaction of buyers and sellers, unlike an open-ended mutual fund where liquidity is offered only at the close and only at net asset value. As such, ETFs can serve as important vehicles for price discovery when the underlying markets are stressed or illiquid. International funds provide daily examples of this point.

Transaction Costs: Externalized

An important difference from a mutual fund structure is that transaction costs in an ETF are "externalized." Consider a hypothetical mutual fund with assets of $100 million and one million shares outstanding. The average bid–ask spreads of the underlying assets are for illustrative purposes assumed to be 0.20 percent, and so one-way transaction costs are 0.10 percent. Suppose on a given day there are $5 million of inflows (subscriptions) and $20 million of outflows (redemptions) for a net outflow of $15 million. Say also that fundamental values remain constant over the day. In the traditional open-ended mutual fund example, subscriptions and redemptions occur at the net asset value of $100, and the fund manager must sell $15 million of the underlying assets. These sales will tend to occur at the bid price of the underlying assets, and hence an average discount of 0.10 percent to net asset value. At the start of the following day, net asset value is—assuming no change in fundamentals—equal to $84,985,000, which is calculated as the original $100 million, minus the $15 million in sold assets, and also minus the transaction costs of selling. In other words, remaining investors in the mutual fund bear the transaction costs incurred by the participants who redeemed or subscribed.

In contrast, in exchange-traded funds, the sellers of the fund will transact directly with buyers at a market determined price. Net selling does not require the ETF manager to interact with the capital markets, meaning that in this example,

fund investors who do not sell will hold a fund whose assets are valued at $85 million.

Moreover, in exchange-traded funds the distribution fees are externalized. In a "compensation model" for financial advisers, which is increasingly common worldwide, financial advisors are paid directly by the client for their services typically based on the amount of assets managed. For these professional advisors, ETFs are attractive because distribution, account servicing, or maintenance fees are not included in the expense ratio. Mutual fund managers often pay financial advisors a commission, called a "retrocession," for selling their products to clients. In Europe, the recent trend towards eliminating these payments (through laws that state that advisors should act in their clients' interests) puts ETFs and mutual funds on par in terms of compensation, from the perspective of a financial advisor. That change should also increase incentives for advisors to offer their clients ETFs as an element of portfolio construction.

Other Considerations

Compared to active mutual funds or to hedge funds, exchange-traded funds offer greater transparency because their investment strategies are specified in advance and their holdings are listed daily versus quarterly. The ETF structure also enables lower fees than traditional active mutual funds. Since mutual funds interact directly with investors (Antoniewicz and Heinrichs 2014; Hill, Nadig, and Hougan 2015) they accrue distribution and record-keeping costs. Indeed, mutual funds may levy fees (such as transfer agency fees or 12b-1 fees that compensate the fund for distribution and service) that ETFs do not, raising the cost to own mutual funds.

An investor in ETF shares, unlike a traditional mutual fund investor, can short shares, lend shares, and can buy on margin, as with stocks. (With short sales, an investor faces the potential for unlimited losses as the security's price rises. There are special risks associated with margin investing. As with stocks, an investor may be called upon to deposit additional cash or securities to their account, there is no guarantee that there will be borrower demand for the ETF, and a short sale may or may not be recalled.)

Relative to open-ended index mutual funds, exchange-traded funds can potentially offer significant tax advantages that derive from the ability to use in-kind transfers to reduce capital gains distributions, as explained in detail in Poterba and Shoven (2002). The ability to trade ETFs intraday also makes them attractive to hedge funds and other institutions seeking to hedge risks or gain exposure based on macroeconomic and other news events.

Potential Issues for ETFs

One potential issue for exchange-traded funds is that some investors may not have the financial sophistication to distinguish between the types of ETFs (for example, funds that are levered or that are based on unsecured debt) in the absence of a common classification scheme. A second issue is that, intraday liquidity might induce "too much" trading. Barber and Odean (2000) show that individual investors who

trade actively in individual stocks suffer lower returns than investors who trade less. The liquidity of ETFs might lead to a similar effect relative to less-liquid mutual funds. Finally, the proliferation of indices, some custom and others concentrated, pose challenges for ordinary investors. Asset managers may create indices that are designed to do well in backtesting but might not do well going forward. We will address potential concerns about the growth of ETFs in more detail later in the paper.

The Size and Types of Exchange-Traded Products

Equity-based exchange-traded funds still dominate the ETF landscape, accounting for over 78 percent of the $4.3 trillion in exchange-traded product assets, but other asset classes (including fixed income, which is 17 percent of assets) have become more important recently (according to BlackRock 2017b).

Distinguishing among different kinds of exchange-traded products is useful given that regulatory concerns about the possible disruptive effects of ETFs often focus on a relatively small subset of the universe of exchange-traded products. For example, *exchange-traded notes* are senior, unsecured (either collateralized or more likely uncollateralized) debt securities that are exposed to the credit risk (solvency) of the issuer, typically an investment bank. Only 2.3 percent of global assets in all exchange-traded products are held in exchange-traded notes. A small subcategory of exchange-traded notes includes ETFs that are not backed by publicly traded holdings; ETFs backed by bank loans are about $7 billion or 0.2 percent of total assets in exchange-traded products. *Exchange-traded commodity funds* are funds that hold physical commodities such as silver or gold. *Leveraged and inverse exchange-traded products,* which represent 1.3 percent of global assets in exchange-traded products, hold the individual index stocks as well and thus have elements of physical-backing (Madhaven 2016).

Table 1 shows the assets under management (AUM) of broad categories of exchange-traded funds, including equity, fixed income, commodity, currency, and alternative/asset allocation ETFs. The vast majority of ETFs, representing 92.5 percent of global assets of nondebt funds are traditional ETFs that typically hold a portfolio of securities (stocks or bonds) that closely resembles, but need not necessarily fully replicate, their benchmark index (Madhavan 2016). These funds seek to provide one-to-one exposure to the index, usually broad market gauges offered by index providers. Beyond helping investors distinguish among exchange-traded products, a sensible classification scheme could help speed up the regulatory process for "plain vanilla" funds comprised of stocks/bonds that do not use leverage, swaps, and other financial tools.

Table 1 also shows the number of different indices tracked by ETFs for a variety of different asset classes. Exchange-traded funds track 130 US large cap indices, the largest ETF sector. In addition to these broad market indices, ETFs seek to track 208 sector indices and hundreds of other more specialized indices. ETFs also span 180 indices across different fixed-income markets as well as 126 commodity and 22 currency indices.

Table 1
ETF Overview

Type of ETF	*Number of distinct benchmarks*	*Assets under management in 2015 ($ millions)*
Equity		
Global equity	92	35,750
US large cap/Total cap	130	383,987
US mid cap	46	59,715
US small cap	56	61,751
US sector	204	158,923
US dividend preferred	23	68,358
US alpha strategy	14	2,109
Developed Europe	36	18,000
Developed Asia Pacific	28	32,202
Emerging/Frontier	158	155,249
International/Other	115	105,418
Fixed income		
Broad market	16	63,687
Emerging markets	11	13,417
High yield	16	32,835
Investment grade	32	60,037
Securitized	4	7,029
Municipals	29	13,690
Sovereign	17	4,867
US Government	55	58,595
Commodities	126	91,865
Currency	22	4,488
Alternatives/Asset allocation	87	8,311

Source: Investment Company Institute (2016).

Equity Exchange-Traded Funds

Table 2 takes a closer look at equity exchange-traded funds.

The growth of ETFs is linked to a broad shift from actively managed mutual funds to passive investment vehicles. During the period from 2007 to 2015, over $425 billion flowed into passive mutual funds and $730 billion into exchange-traded funds, while actively managed mutual funds lost $835 billion in assets under management (Investment Company Institute 2016). It is also worth noting that until the advent of electronic data delivery and cheaper computing technology, it was quite costly to manage an index portfolio of several hundred or thousand constituents relative to a concentrated active portfolio of, say, 50–70 stocks. Indeed, it was only in the 1970s that it became cost effective to manage an index fund. ETFs succeeded in the 1990s as a result of regulation that saw them as a way to provide market stability after the crash of 1987 without portfolio trading of individual stocks (as reported in Balchunas 2016).

Yet despite the shift into index vehicles, considerable room for growth remains. The global investable universe for equities—the value of all publicly traded company stocks—is an estimated $68 trillion (according to BlackRock 2017a). Traditional open-end mutual funds, index and active, hold approximately

Table 2
Equity ETF Types

Type of ETF	*Assets under management in 2015 ($ millions)*
Market cap based	1,007,059
Total market	446,615
Large cap	414,979
Mid cap	70,935
Small cap	74,529
Sector	273,753
Factor/Smart beta	435,701
Growth/Value	230,529
Dividend	92,367
Equal weight	28,918
Low volatility	23,810
Multi factor	42,246
Single factor	17,830
Momentum	3,840
Quality	2,474
Value	2,068
Size	4,463
Other	4,985
Other	11,527
Total	**1,728,040**

Source: Investment Company Institute (2016).

15.2 percent and 4 percent, respectively, of the investable equity universe. (Among open-end mutual funds, index funds represent 7.4 percent of the equity universe.)

Fee differentials and the difficulties of beating a benchmark may explain some of the movement from active to passive indexing, including exchange-traded funds. The management fees for mutual funds have declined in recent years: in 2000, management fees of active mutual funds on average were 106 basis points, about 80 basis points higher than fees of index mutual funds. By 2015, average fees of active funds declined by about 20 basis points while average fees of index funds have declined by 16 basis points (Investment Company Institute 2016). Average fees of bond mutual funds have declined by a comparable margin. The fees for exchange-traded funds are typically lower than actively traded mutual funds but higher than those for passive index mutual funds. The majority of mutual funds have not outperformed their benchmarks once fees are taken into account (for example, Carhart 1997; Grinblatt and Titman 1992; Elton, Gruber, and Blake 2011).

In Table 2, the second category of equity ETFs (after the market-cap-based ETFs) is the sector exchange-traded funds, which typically seek to track market-weighted capitalization benchmarks for each sector. The main sectors that are represented by ETFs, each with about $10–$13 billion in assets under management, are (from larger to smaller) natural resources, real estate, financial services, health, technology, and consumer goods. It is interesting to note that the shares of these

specific sectoral funds among the total for all sectoral funds are similar to the corresponding sector weights in the S&P 500 index.

The third category of equity ETFs on Table 2 is so-called "smart beta" or factor exchange-traded funds. These ETFs follow weighting schemes that differ from traditional market cap-based indices and are primarily driven by the desire to outperform the market portfolio by focusing on certain factors that have been linked to stock returns. Smart beta ETFs blur the lines between traditional active versus passive investment strategies. On the one hand, these ETFs offer exposure to risk factors that traditionally have been exploited by active mutual funds and hedge funds. On the other hand, smart beta ETFs track specific indices in a transparent and rule-based manner, and there is no active money manager who "picks" stocks. Consequently, the expense ratios of factor ETFs are typically lower than those of comparable active mutual funds and hedge funds. These ETFs have become more popular recently, but as Table 2 shows, factor/smart beta ETFs accounted for about 25 percent of total equity assets under management. The importance of factor/smart beta ETFs is expected to grow as investors seek to capture factor premia.

What are some of the common factors that smart beta funds seek to capture? The largest factor ETF category focuses on "value stocks" and "growth stocks," a categorization that goes back to Graham and Dodd (1934). Growth stocks tend to have high ratios of stock prices to fundamentals, such as earnings, sales, and book values. In contrast, value stocks have low price-to-earnings and high book-to-market ratios. A large academic literature has investigated the risk and returns of value and growth stocks going back to Ball (1978) and Basu (1983), as summarized by Ang (2014) and Bali, Engle, and Murray (2016). The key finding is that value stocks have outperformed growth stocks, and this "value premium" cannot be explained by traditional risk models, such as the classic single-factor Capital Asset Pricing Model. Before the advent of factor ETFs, investors had two options to gain exposure to value/growth stocks: either they had to purchase individual stocks directly from a broker or they invested in actively managed value/growth mutual funds. Both options carry significant transaction costs and/or management fees. Factor ETFs enable investors a similar objective at significantly lower cost.

While growth/value ETFs represent by far the largest fraction of factor ETFs, many ETFs track other "factors" such as dividend yield or momentum. For example, long–short factors discussed in Fama and French (2015) include: "high minus low," which is a long–short portfolio that invests in high book-to-market value stocks and shorts high book-to-market growth stocks; "small minus big," which is long in small stocks and short in large stocks; "up minus down," which is a momentum factor that is long in stocks that have had high return over the previous year and short in stocks that had low returns; "robust minus weak," which is the difference between returns of profitable firms and unprofitable firms; and "conservative minus aggressive," which is the difference between returns of firms that invest a lot and firms with low investment rates. Unlike the long/short factors used in academic research, most ETF factor funds are long-only. Factor ETFs are low-cost investment vehicles for investors who seek long-only exposure to well-known factor risks with lower fees than active

mutual fund and hedge fund managers.[3] Some recent "smart beta" ETFs combine multiple factors to exploit diversification and correlations across factors, and seek exposure to risk premia (for example, exchange-rate risk) beyond just equities.

In 2008, the Securities and Exchange Commission adopted new guidelines for listing of active ETFs. These ETFs have a benchmark index, as passive ETFs, but allow the ETF manager discretionary portfolio decisions with the goal of outperforming the benchmark. Unlike active mutual funds and hedge funds, active ETFs are required to disclose their portfolio holdings daily. Active ETFs, while still a fraction of total assets, further blur the lines between active and passive investment management. The complexity of mutual funds and ETFs requires careful research and financial sophistication on the part of potential investors.

Fixed Income and Commodity Exchange-Traded Funds

Fixed income exchange-traded funds (going back to Figure 1) have grown dramatically in recent years. Initially, these were typically portfolios of investment grade and government bonds; more recently, bond ETFs have been created based on high-yield bonds and even bank loans. As of September 2017, bond ETFs account for about 17 percent, or $740 billion, of total assets invested in ETFs.

What explains this rapid growth in bond exchange-traded funds? Investors in individual bonds face a number of challenges. First, many corporate bonds are traded primarily in the opaque, dealer ("over-the-counter") market. By contrast, bond EFTs trade intraday on electronic exchanges, many with low bid–offer spreads compared to the underlying bonds (for example, Hendershott and Madhavan 2015). Second, unlike individual bonds, fixed income ETFs offer a high degree of transparency, meaning that bid and offer quotes are readily available. Third, many individual bonds are illiquid and trade infrequently. Bid–ask spreads in bond markets can be significantly higher than spreads in equity markets, while exchange-traded bond funds typically offer greater liquidity and diversification. Fourth, keeping the maturity of a bond portfolio constant requires constant trading, but a bond ETF can be designed to do this without the need for ongoing attention and trading.

Bond exchange-traded funds are attractive to individual bond buyers—either retail or institutional—in the context of these challenges. Pension funds have started to embrace the concept of passive investing in fixed income assets because of low cost, diversification, and transparency. Other investor types, such as hedge funds or large institutions, may use bond ETFs as exposure vehicles or ways to invest cash.

There has also been considerable interest in commodity-based exchange-traded funds, often viewed as a hedge against inflation or a source of diversification, although the role of commodity ETFs has declined since 2013 when prices of many commodities fell dramatically. Commodity ETFs for the most part must invest indirectly via futures contracts, with the exception of certain precious metals (including

[3] In an online Appendix available with this article at http://e-jep.org, we offer some sample calculations of how the returns to actively managed mutual funds compare with the returns from a portfolio based on these kinds of factors, along with sector funds.

gold), because the physical costs of storage of commodities would push the expenses of a commodity ETF far too high (Madhavan 2016). Because ETF commodity funds offer exposure via futures contracts (including those on esoteric asset classes such as volatility), they need not always reflect spot returns.[4]

Concerns and Misconceptions

An investor can lose money with exchange-traded funds, of course, just as an investor can lose money with mutual funds, hedge funds, or any of the underlying assets. The salient question here is whether there may be certain kinds of risks with exchange-traded funds that make them riskier than commonly perceived—either for individual investors, or for financial markets, or even for the economic system as a whole. We will argue that while certain concerns do exist with regard to ETFs, as they do for other financial markets, the concerns are often based on misconceptions. We begin with concerns for individuals and then move to questions of the broader impact of index investing on the markets and the macroeconomy.

Fund Closures, Shorting, and Counterparty Risk

Individual investors often worry about the risk of losing their entire investment. Closures of exchange-traded funds, like the closures of mutual funds, are not uncommon. Anywhere from 50 to 80 exchange-traded funds close each year (Madhavan 2016).

While the closure of an ETF can attract attention, it does not create investment risk in itself (unlike a firm's bankruptcy), as the fund's underlying assets should not be affected. When an ETF closes, its price should converge to its net asset value. A plain-vanilla unlevered fund is just a pool of assets, and should the fund be redeemed in full, the assets can potentially simply be returned in kind. Of course, investors in a fund to be closed may experience unanticipated capital gains taxes and, for a time, a possible lack of liquidity.

For other exchange-traded products, these risks may be greater. In 2008, Lehman Brothers had issued exchange-traded notes that were unsecured debt obligations. When Lehman Brothers declared bankruptcy, there were no underlying assets to be returned to investors. This case highlights our earlier remarks regarding the need for a classification scheme to help investors distinguish between the various types of exchange-traded products. There can also be counterparty risk, when certain synthetic exchange-traded funds enter into swap positions with investment banks. However, the risk that any given counterparty might fail is mitigated by diversification rules that spread the risk across multiple swap counterparts. It

[4] Madhavan (2016) shows the impact of the futures forward curve for volatility, where the normal upward slope of the curve implies negative returns on average to an investor who rolls from near to far contracts to gain exposure to spot volatility.

is unlikely that such losses could exceed the assets of an exchange-traded fund, because even a leveraged fund is collateralized with cash and securities.

Let us turn now from fund closures to other concerns that could lead to significant individual investor losses, and possibly larger impacts on the financial system. Specifically, one possible concern is that when exchange-traded funds are sold short, the aggregate long and synthetic long positions can exceed the total actual number of outstanding ETF shares (for example, Bradley and Litan 2010). If many investors simultaneously redeem their shares in an ETF at the same time, some argue that this could theoretically "bankrupt" the fund, as redemptions would exceed available assets to be redeemed. However, institutional details around ETF settlement make this scenario remote. On the settlement day, ETF managers only release redemption proceeds against actual delivery of the ETF shares. An *attempt* to redeem by a party that does not actually physically have ETF shares to deliver (say, because they have lent their shares to a short seller) will simply fail to settle. It is possible that the failure of a large number of such attempted "redemptions" could itself result in market disruption, but this scenario seems remote.

A closely related set of concerns involves securities lending and counterparty risk. Securities lending is the temporary transfer of a security by its owner (for example, a pension fund) to another party (for example, a hedge fund), typically for the purposes of a short-sale. The lender remains the owner of the security, and hence is exposed to any security price movement over the life of the loan. The borrower usually provides collateral (typically in excess of the security's value ranging from 102–112 percent) to compensate the lender in the rare case that the borrower fails to return the borrowed security.

Can securities lending by an exchange-traded fund pose a threat to investors? First note that in the United States there is presently a 50 percent aggregate statutory limit on the extent to which exchange-traded funds can lend their underlying securities. Moreover, other safeguards on lending include the ability to recall loans from borrowers and possibly even the liquidation of the borrower's collateral. Securities lending may help enhance ETF returns when safeguarded in these ways. From a market perspective, securities lending can help improve liquidity and price efficiency by reducing the costs of expressing negative views through short-selling, helping to keeps asset bubbles from forming. Although securities lending is prevalent and economically significant, the academic literature on securities lending is nascent.

Flash Events and Systemic Risk

Another issue that concerns both individuals and regulators concerned with the broader markets are "flash events," marked by sharp price movements and subsequent reversals in compressed time intervals). In the "Flash Crash" of May 6, 2010, the Dow Jones Industrial Average dropped almost 1,000 points in 20 minutes. Many well-known stocks briefly traded at clearly unreasonable prices, including some that traded at pennies.

Exchange-traded funds were disproportionately represented among the securities most affected (for discussion, see Borkovec, Domowitz, Serbin, and Yergerman

2010), with prices diverging widely from their underlying net asset values, which led some commentators to draw a connection from the sharp market moves on May 6 to the pricing and trading of these instruments (for example, Wurgler 2011).[5]

Madhavan (2012) also describes some market structure issues, including increased market fragmentation and the proliferation of new venues, which could be factors in a flash event. He also finds evidence that aggressive "order-sweeping" trades—that is, a large trade executed all at once at whatever range of prices are being offered at the moment, rather than spread out over time in an attempt to get the best possible price—were related to the market dislocation, as opposed to structural problems with ETFs. A similar flash event in August 2015 has led many industry participants, including asset managers, brokers, and exchanges, to organize and implement many important changes to market structure.

Flash events have taken place in other asset classes since 2010, including US Treasury bonds and currencies, where ETFs are minor. On October 15, 2014, the yield on the 10-year US Treasury note fell to 1.86 percent before reversing to 2.13 percent within a 15-minute time interval. A Joint Staff Report (US Department of the Treasury et al. 2015) by staff of US Treasury, the Federal Reserve, and financial regulators found that the intraday yield change was eight standard deviations greater than normal and noted: "For such significant volatility and a large round-trip in prices to occur in so short a time with no obvious catalyst is unprecedented in the recent history of the Treasury market." This report found that speed and size of the yield changes seems to trace back to the evolving structure of the Treasury market, including the role of automated trading. As another example, the value of the UK pound sterling dropped by more than 6 percent against the US dollar in just a few minutes on October 6, 2016, falling to a record low of $1.1378 (as reported in McDonald 2016). These recent flash events highlight that the need for further research on liquidity gaps in increasingly fast markets.

Liquidity Mismatch

Liquidity is often described as the ability to buy or sell without causing substantial price changes. In the case of exchange-traded funds, liquidity concerns can arise at several levels. Liquidity in the *primary* market, where the underlying securities trade, refers to the ability of Authorized Participants to acquire the underlying assets and transfer them in-kind (or vice versa) to the ETF provider for shares in the fund or vice versa. The key role of Authorized Participants in adjusting the ETFs shares outstanding to reflect supply and demand has often given rise to questions of systemic risk if they should "step away" in a crisis. But if a particular Authorized Participant ceased its activities in a certain ETF, other Authorized Participants seem highly likely to provide liquidity. A comprehensive analysis of 931 US exchange-traded funds covering $1.8 trillion of assets under management by the Investment

[5] Ramaswamy (2011) examines the operational frameworks of exchange-traded funds and relates these to potential systemic risks. The role of leveraged ETFs has also been discussed (for example, Cheng and Madhavan 2009) in the context of end-of-day volatility effects.

Company Institute (Antoniewicz and Heinrichs 2015) shows that the largest ETFs—those of most concern from a systemic risk viewpoint—have an average of 38 Authorized Participants. These issues are unlikely to be a concern for ETFs with many Authorized Participants (which is most ETFs) since it is an unlikely event that all Authorized Participants jointly cease their activities at the same time, but may be relevant for smaller niche ETFs with just a few Authorized Participants. If all Authorized Participants were to withdraw, the ETF would likely trade like a closed-end mutual fund (that is, a fund with a fixed number of shares) with possibly wider premiums or discounts.

A second set of concerns relate to the so-called secondary markets, the venues where shares of exchange-traded funds actually trade. The liquidity (measured by dollar volume) in the secondary market can be many times that of the primary market, as discussed earlier. In that sense, the ETF liquidity in the secondary market (via the creation/redemption mechanism of arbitrage) is generally greater than or equal to the liquidity of the underlying assets. The trading of ETF shares on exchanges in the secondary market does not directly drive buying and selling of the underlying stocks but rather reflects changes of ownership of the ETF. Purchases and sales of stocks driven by the ETF creation and redemption process account for only 5 percent of all US stock market trading. In other words, the existence of ETFs can add a layer of incremental liquidity to the financial markets. From a financial stability viewpoint, this buffer is additive.

Impact on Underlying Markets

Some commentators have raised questions about the effect of index investing—including index mutual funds and exchange-traded funds—as a potential distortion of the prices of underlying securities. From an academic perspective, the implications of the introduction of a "basket" security like a diversified index mutual fund or ETF are not clear. Individual investors can reduce their own costs of trading with informed agents by using basket securities as their asymmetric information costs will be lower (Kyle 1985). To the extent that "noise traders" migrate to the basket market, liquidity in the underlying stocks or bonds may decline. However, the creation of a low-cost diversified basket instrument may also open up access to new liquidity investors who were previously unable to access the market due to cost or other constraints. This means that the impact of a basket security on liquidity of the underlying market bonds is an empirical question (for arguments that ETF trading adds additional volatility, see Dannhauser 2017; Ben David, Franzoni, and Moussawi 2017).

But in practical terms, the relative scale of index investing is still relatively small. Index investing overall represents less than 20 percent of global equities (BlackRock 2017b). Index funds and ETFs together represent just over 12 percent of the US equity universe, and 7 percent of the global equity universe. Also, focusing on the dollar size of indexed assets diverts attention from the real issue, namely the turnover by fund managers. Specifically, if we look more closely at US equities, the majority of the assets in funds are actively managed, and active fund managers have significantly greater turnover than passive index funds or ETFs.

As previously noted, there is general agreement on the private benefits of indexing as an efficient way to invest in lieu of paying for security selection. Questions and concerns have increasingly shifted to the impact of index investments on pricing in financial markets (that is, social impact), and some commentators have suggested that the growth of indexing can cause prices to decouple from value. Index trackers are typically based on market capitalization weighted schemes, so some argue that pricing errors in underlying stocks might feed on themselves; a bubble in, say, tech stocks is reinforced by the mechanical action of index funds who are price takers. Could ETF flows distort prices? Index funds are price-takers, not price-makers. They invest, proportionally at whatever price is determined by the buying and selling of active participants. So index assets are a proportional slice of the overall market—that is, a slice of the aggregate value of all securities. The value of all active and other, non-indexed assets is just the overall market less all index assets. Therefore, the money coming into index funds/ETFs must come from the pool of non-indexed/active assets, which (from above) is a slice that is proportional to the overall market, at all points in time.

For index flows to distort prices, one would have to argue that despite having an origin in a pool proportional to the overall market, the desire for index exposures is manifested very differently in characteristics such as capitalization, sector, and so on. While this is possible, there is no evidence that this is true. What about smart beta and other tilts that systematically deviate from capitalization weights? They are still tiny relative to the overall market (Ang, Madhavan, and Sobczyk 2017).

Now consider the arguments about the impact of index inclusion on return correlations and comovement of stocks. As many studies have shown, the average pairwise return correlation between any two stocks has increased since 2000, a period of rapid growth in ETF and index assets, but this trend followed a dramatic decline in pairwise correlations from the 1970s to the late 1990s (Campbell, Lettau, Malkiel, and Xu 2001). Moreover, cross-stock correlations were higher in the 1930s before the advent of indexing (Madhavan 2016). Comovements among currencies—an area with no meaningful index penetration—have similarly risen in the past decade, again a reflection of the importance of central bank policy and a macro-driven environment. Correlations have diminished significantly since 2013 despite significant increases in ETF and index assets (as of March 2017).

The success of active fund management has more to do with the dispersion of returns than correlations. When common factors explain a large fraction of return movements relative to security-specific return, correlations will by definition be large, and the opportunities for professional managers will be correspondingly lower. Moreover, active bets are zero sum irrespective of the correlation environment. That is not to say that active managers cannot profit from active bets by other investors who may hold active positions for behavioral or other reasons (like tax reasons or desire for stock in a certain company). Our point is that the share of active and passive management is determined in a self-regulating manner. Markets will reach an equilibrium when security selectors as a group break even after taxes and fees (Berk and Green 2004; Pástor, Stambaugh, and Taylor 2015).

Conclusions

Exchange-traded products provide exposure to a wide range of asset classes (for example, equities, fixed income commodities, and currencies), strategies (for example, passive index, model-based, and active), and regions. Exchange-traded funds have grown substantially in diversity and size in recent years along with the rise of passive, index investing. Equities still account for over 78 percent of assets under management in ETFs as of 2017 (but there is rapid growth in all asset classes, and fixed income in particular, with assets now in excess of $740 billion or 17 percent of the total in all exchange-traded products (according to *BlackRock* 2017b).

The discussion in this paper has suggested a number of reasons behind this growth. First, there are the traditional advantages of exchange-traded funds in terms of liquidity, low fees, transparency, and potential tax advantages. Second, the universe of ETFs has been expanding beyond the traditional equity-based funds, including funds providing access to fixed income, commodities, currency, volatility, multi-asset class structures, and "smart beta" or factors. Many of these new ETFs represent a blurring of the traditional line between active and passive management. Third, the investor base of ETFs has also been expanding. As bank balance sheets shrink in the new regulatory environment after the 2008–2009 financial crisis, ETFs are being used by institutional investors as a substitute for futures, credit derivatives, swaps, and individual bond trading. Professional financial advisors and hedge funds are making greater use of ETFs in a number of ways. Model portfolios using ETFs and the rise of robo-advisors are also longer-term trends that favor ETF use and adoption.

There is little evidence of pressures or flaws that have uniquely affected ETFs compared to other equity investment vehicles. Is turnover excessive? Do ETFs encourage overtrading? These are valid questions that also arise for other low-cost vehicles for broad market price discovery such as futures or swaps. Indeed, US futures trade approximately $250 billion a day, with a high concentration of volume in S&P 500 and Russell 2000 portfolios; by contrast, ETFs are traded in far more diverse portfolios including domestic and international equity, commodities, fixed income, and alternatives. Moreover, the advent of discount brokerages has dramatically reduced the cost of participating in financial markets. While such decreasing cost of trading can be a double-edged sword (allowing broader participation in financial markets while encouraging excessive trading), there is no evidence that financial markets have become less efficient. In modern well-developed financial markets there are many vehicles for correcting mispricing at the individual security level—for example, trading by individuals or sovereign wealth funds, along with share repurchase and issuance, trading of stock options, and the ability to take companies private/public.

This paper also surveyed potential concerns for individuals as well as markets as a whole, echoing the increased scrutiny of exchange-traded funds in the media and by regulators. A problematic aspect of this discussion is that not enough attention is paid to the diversity of the ETF landscape. There is no single "ETF." Instead, potential concerns apply to some ETF types but not to others. The vast majority of

market share is invested in traditional passive, unlevered, cap-based ETFs, which share many features of index mutual funds. It seems important to take a more nuanced view that distinguishes the various ETF types in the same way we assess the pros and cons of mutual fund types differently. From the perspective of an individual investor, the increased variety and complexity of investment options, while providing more opportunities, requires more financial sophistication. ETFs are part of this trend with advantages and possible disadvantages.

■ *The views expressed here are those of the authors alone and do not necessarily represent the views of BlackRock, Inc.*

References

Ang, Andrew. 2014. *Asset Management: A Systematic Approach to Factor Investing.* Oxford University Press.

Ang, Andrew, Ananth Madhavan, and Aleksander Sobczyk. 2017. "Crowding, Capacity, and Valuation of Minimum Volatility Strategies." *Journal of Index Investing* 7(4): 41–50.

Antoniewicz, Rochelle, and Jane Heinrichs. 2014. "Understanding Exchange-Traded Funds: How ETFs Work." *ICI Research Perspective*, September, vol. 20, no. 5, Investment Company Institute, Washington, DC. https://www.ici.org/pdf/per20-05.pdf.

Antoniewicz, Rochelle, and Jane Heinrichs. 2015. "The Role and Activities of Authorized Participants of Exchange-Traded Funds." March. Investment Company Institute, Washington, DC. http://www.ici.org/pdf/ppr_15_aps_etfs.pdf.

Balchunas, Eric. 2016. "The ETF Files: How the U.S. Government Inadvertently Launched a $3 Trillion Industry." *Bloomberg Markets,* March 7. https://www.bloomberg.com/features/2016-etf-files/.

Bali, Turan G., Robert F. Engle, and Scott Murray. 2016. "Empirical Asset Pricing: The Cross Section of Stock Returns." Wiley.

Ball, Ray. 1978. "Anomalies in Relationships between Securities' Yields and Yield-Surrogates." *Journal of Financial Economics* 6(2–3): 103–26

Barber, Brad M., and Terrance Odean. 2000. "Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors." *Journal of Finance* 55(2): 773–806.

Basu, Sanjoy. 1983. "The Relationship between Earnings' Yield, Market Value and Return for NYSE Common Stocks: Further Evidence." *Journal of Financial Economics* 12(1): 129–56.

Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi. 2017. "Exchange-Traded Funds." *Annual Review of Financial Economics* 9: 169–89.

Berk, Jonathan B., and Richard C. Green. 2004. "Mutual Fund Flows and Performance in Rational Markets." *Journal of Political Economy* 112(6): 1269–95.

Berk, Jonathan B., and Richard Stanton. 2007. "Managerial Ability, Compensation, and the Closed-End Fund Discount." *Journal of Finance* 62(2): 529–56.

BlackRock. 2017a. "Index Investing Supports Vibrant Capital Markets." ViewPoint, September. https://www.blackrock.com/corporate/en-us/literature/whitepaper/viewpoint-index-investing-supports-vibrant-capital-markets-oct-2017.pdf.

BlackRock. 2017b. "BlackRock Global ETP Landscape: Monthly Snapshot, October 2017." https://www.blackrock.com/au/intermediaries/literature/market-commentary/global-etp-landscape-en-aus.pdf.

Borkovec, Milan, Ian Domowitz, Vitaly Serbin, and Henry Yegerman. 2010. "Liquidity and Price Discovery in Exchange-traded Funds: One of Several Possible Lessons from the Flash Crash." *Journal of Index Investing* 1(2): 24–42.

Bradley, Harold, and Robert E. Litan. 2010. *Choking the Recovery: Why New Growth Companies Aren't Going Public and Unrecognized Risks of Future Market Disruptions.* Kauffman Foundation.

Campbell, John Y., Martin Lettau, Burton G. Malkiel, and Yexiao Xu. 2001. "Have Individual Stocks Become More Volatile? An Empirical

Exploration of Idiosyncratic Risk." *Journal of Finance* 56(1): 1–43.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52(1): 57–82.

Cheng, Minder, and Ananth Madhavan. 2009. "The Dynamics of Leveraged and Inverse Exchange-Traded Funds." *Journal of Investment Management*, Fourth Quarter, 7(4). (Also available at SSRN: https://ssrn.com/abstract=1539120.)

Dannhauser, Caitlin D. 2017. "The Impact of Innovation: Evidence from Corporate Bond Exchange-Traded Funds (ETFs)." *Journal of Financial Economics* 125(3): 537–60.

Elton, Edwin J., and Martin J. Gruber. 2013. "Mutual Funds." Chap. 15 in *Financial Markets and Asset Pricing: Handbook of Economics and Finance*, Vol. 2 (Part B), edited by George M. Constantinides, Milton Harris, and Rene M. Stultz. Elsevier.

Elton, Edwin J., Martin J. Gruber, and Christopher R. Blake. 2011. "Holdings Data, Security Returns, and the Selection of Superior Mutual Funds." *Journal of Financial and Quantitative Analysis* 46(2): 341–67.

Elton, Edwin J., Martin J. Gruber, and Jeffrey A. Busse. 1998. "Do Investors Care about Sentiment?" *Journal of Business* 71(4): 477–500.

Fama, Eugene F., and Kenneth R. French. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116(1): 1–22.

Graham, Benjamin, and David L. Dodd. 1934. *Security Analysis*. New York: McGraw-Hill Book Company.

Grégoire, Vincent. 2013. "Do Mutual Fund Managers Adjust NAV for Stale Prices?" Available at SSRN: https://ssrn.com/abstract=1928321.

Grinblatt, Mark, and Sheridan Titman. 1992. "The Persistence of Mutual Fund Performance." *Journal of Finance* 47(5): 1977–84.

Hendershott, Terrence, and Ananth Madhavan. 2015. "Click or Call? Auction versus Search in the Over-the-Counter Market." *Journal of Finance* 70(1): 419–47.

Hill, Joanne M., Dave Nadig, and Matt Hougan. 2015. "A Comprehensive Guide to Exchange-Traded Funds." CFA Institute Research Foundation.

Investment Company Institute. 2016. *2015 Investment Company Fact Book*. https://www.ici.org/pdf/2015_factbook.pdf.

Kyle, Albert S. 1985. "Continuous Auctions and Insider Trading." *Econometrica* 53(6): 1315–36.

Lee, Charles M. C., Andrei Shleifer, and Richard H. Thaler. 1990. "Anomalies: Closed-End Mutual Funds." *Journal of Economic Perspectives* 4(4): 153–64.

Lee, Charles M. C., Andrei Shleifer, and Richard H. Thaler. 1991. "Investor Sentiment and the Closed-End Fund Puzzle." *Journal of Finance* 46(1): 75–109.

Madhavan, Ananth. 2012. "Exchange-Traded Funds, Market Structure, and the Flash Crash." *Financial Analysts Journal* 68(3): 20–35.

Madhavan, Ananth. 2016. *Exchange-Traded Funds and the New Dynamics of Investing*. Oxford University Press: New York, NY.

Madhavan, Ananth, and Aleksander Sobcyzk. 2016. "Price Dynamics and Liquidity of Exchange-Traded Funds." *Journal of Investment Management*, forthcoming.

McDonald, Sarah. 2016. "Pound Is the Latest Flash Crash That Traders Won't Easily Forget." Bloomberg, October 7. http://www.bloomberg.com/news/articles/2016-10-07/pound-is-the-latest-flash-crash-that-traders-won-t-easily-forget.

Pástor, Ľuboš, Robert F. Stambaugh, and Lucian A. Taylor. 2015. "Scale and Skill in Active Management." *Journal of Financial Economics* 116(1): 23–45.

Poterba, James, M., and John B. Shoven. 2002. "Exchange-Traded Funds: A New Investment Option for Taxable Investors." *American Economic Review* 92(2): 422–27.

Ramaswamy, Srichander. 2011. "Market Structures and Systemic Risks of Exchange-Traded Funds." BIS Working Paper 343, Bank of International Settlements. April.

US Department of the Treasury, Board of Governors of the Federal Reserve System, Federal Reserve Bank of New York, US Securities and Exchange Commission, and US Commodity Futures Trading Commission. 2015. *Joint Staff Report: The US Treasury Market on October 15, 2014*. (Joint Staff Report 2015 (July 13).) https://www.treasury.gov/press-center/press-releases/Documents/Joint_Staff_Report_Treasury_10-15-2015.pdf.

Wurgler, Jeffrey. 2011. "On the Economic Consequences of Index-Linked Investing." Chap. 3 in *Challenges to Business in the Twenty-First Century*, edited by Gerald Rosenfeld, Jay W. Lorsch, and Rakesh Khurana. American Academy of Arts and Sciences.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 155–178

Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care?

Benjamin Handel and Joshua Schwartzstein

In a number of situations, there is strong evidence that people do not translate readily available information into the knowledge that would help them make better decisions. For example, people may choose a health insurance plan that costs $500 per year more in premiums in order to obtain a deductible that is $250 lower—despite having access to open enrollment booklets containing relevant information (Handel 2013; Bhargava, Loewenstein, and Sydnor 2017). People buy branded drugs over equivalent but less-expensive generics (Bronnenberg, Dubé, Gentzkow, and Shapiro 2015) even though information printed on the package reveals their equivalence. Investors pay a range of fees for investing in S&P 500 index funds—and index funds with higher fees have meaningful market shares (Hortaçsu and Syverson 2004). Consumers appear to demand the wrong cell phone plans given their previous usage patterns (Grubb and Osborne 2015).

Why don't people use available information? The many possibilities discussed in the research literature broadly fall into two camps, which we refer to as *frictions* and *mental gaps*. The frictions camp focuses on costs of acquiring and processing information. A consumer shopping in a health insurance exchange incurs a cost to explore more of the options in the choice set and to assess them. This camp, and the closely related framework of "rational inattention," maintains the neoclassical assumption that people form accurate beliefs using the information that is worth

■ *Benjamin Handel is Associate Professor of Economics, University of California, Berkeley, California. Joshua Schwartzstein is Assistant Professor of Business Administration, Harvard Business School, Boston, Massachusetts. Their email addresses are handel@berkeley.edu and jschwartzstein@hbs.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.1.155

doi=10.1257/jep.32.1.155

processing, but it incorporates realistic assumptions on how paying attention to or processing information is costly (Stigler 1961; McCall 1970; Caplin and Dean 2015; Sims 2003; Woodford 2012; Gabaix 2014).

The second camp deals with "mental gaps" or psychological distortions in information-gathering, attention, and processing. A consumer in the insurance exchange may neglect important information in selecting plans even if this information is readily available, perhaps from using an incorrect model, (for example, Schwartzstein 2014) or overweighting salient plan features (for example, Bordalo, Gennaioli, and Shleifer 2012, 2013; Kőszegi and Szeidl 2012). This camp emphasizes how, for a variety of reasons, there is a gap between what people think and what they should rationally think given costs. The categories of frictions and mental gaps are not mutally exclusive or exhaustive, but are intended as a broad classification of approaches that researchers take to studying poorly informed choice.

Most empirical research on frictions or mental gaps assumes that one mechanism dominates, without explicit consideration of possible alternatives, or doesn't try to specify the precise underlying mechanism. A primary reason is that, even with extensive data, it can be very difficult in a number of contexts to identify the source of apparent mistakes. For example, a researcher who observes that consumers of health insurance fail to switch to more valuable options over time may have a hard time distinguishing possible explanations including 1) high time costs of search and switching or 2) incorrect views of how likely product values are to change over time. When researchers assume that one specific mechanism underlies poorly informed choices, but cannot credibly distinguish between that mechanism and others, spurious conclusions often follow.

Beyond specifying the extent of and reason for poorly informed choices, a further goal of the literature is to investigate the consumer welfare (henceforth "welfare") impacts of policies in environments where consumers make such choices. When is it important for policy assessments to distinguish between the underlying mechanisms? We define two classes of policies. An *allocation policy* directly allocates (or strongly steers) consumers to specific actions. To assess the welfare impact of an allocation policy, it is sufficient to identify the combined effect of frictions and mental gaps empirically. A *mechanism policy* instead targets specific mechanisms, where policy predictions depend critically on understanding relative magnitudes of different frictions and mental gaps.[1] This classification may require some judgment to apply, but is intended to highlight factors to keep in mind. Our discussion largely focuses on *counterfactual policies*, by which we mean policies that are hard to evaluate empirically before implementing them, but we will also touch on the case of policies that can more easily be studied in action.

[1] The contrast between mechanism and allocation policies is not the same as the distinction between nudges and traditional policy instruments as introduced by Thaler and Sunstein (2009) and analyzed in detail by Farhi and Gabaix (2017). Many nudges, such as reminders, could be viewed as mechanism policies that target "behavioral" mechanisms, like forgetfulness, but we will view others, such as defaults, as allocation policies.

We begin by describing evidence across contexts in which consumers are not using important information. We then outline key frictions and mental gaps that could matter in these contexts. While many empirical papers describe their findings as related to one specific friction or mental gap, they typically provide little evidence to distinguish between mechanisms. After spelling out this key issue, we turn to three related questions. First, what can we say about the magnitude of frictions and mental gaps when we are uncertain about the mechanism? Second, how can we empirically distinguish the mechanisms? Third, for which policy questions is it sufficient to understand magnitudes and for which is it important to distinguish mechanisms?

Some Examples of Information that People Do Not Use

There is a substantial body of research documenting situations and consequences of people not using readily available information. Table 1 provides examples from the domain of health, and Table 2 provides a broader set of examples. Most of these papers do not attempt to distinguish, explicitly or implicitly, between reasons for not using information.

Consumer Ignorance and Misinformation in Health Markets

Consider a scenario: You have a headache, go to a pharmacy, and choose Advil over store-branded medication containing ibuprofen—which is the same active ingredient contained in Advil. This type of choice is common. Bronnenberg et al. (2015) find that the average consumer chooses national headache-remedy brands over chemically equivalent store-brand alternatives 26 percent of the time. What's going on? At a broad level, consumer misinformation appears to be a factor. Bronnenberg et al. find that pharmacists choose national headache-remedy brands over store-brand alternatives only 9 percent of the time, and nonexpert consumers are presumably less knowledgeable about active ingredients and relative safety. A subset of Nielson panelists were asked to name the active ingredient in national headache remedies. The average respondent answered 59 percent of these questions correctly, compared to over 85 percent for nurses, pharmacists, and doctors. Having this knowledge is highly positively associated with purchasing the store brand, as is reporting a belief that store brands are "just as safe" as national brands. This evidence strongly suggests that a lack of knowledge contributes to nonexpert consumers' demand for national brands. But the evidence has less to say about why consumers are misinformed.

Other papers documenting mistakes in the health treatment decisions of consumers likewise do not typically attempt to identify the causes or domains of misinformation. Pauly and Blavin (2008) and Baicker, Mullainathan, and Schwartzstein (2015) summarize evidence that people have a systematic propensity to under- or overuse certain treatments at the margin. For example, Choudhry et al. (2011) document that many recent heart attack victims do not adhere to drug

Table 1
Examples of Information People Don't Use in Health Markets

Paper	*Findings*	*Potential explanations for not using information*
Health insurance		
Handel and Kolstad (2015b)	"Uninformed" consumers leave substantial dollars on table when "over-choosing" generous insurance coverage, relative to "informed" consumers. Consumers who think (incorrectly) that more generous coverage gives them access to generous providers are willing to pay much more (~$2,300) for that coverage.	*Frictions:* Search costs lead to limited information; information processing costs lead to poor evaluations of plan characteristics. *Mental gaps:* Mistaken beliefs about important ways plans differ; neglect of key plan characteristics.
Bhargava, Loewenstein, and Sydnor (2017)	In active choices, consumers frequently choose dominated plans from menu of 48 insurance options at large employer, losing $300–$400 on average. Experiments show better choices in simplified choice environments and in environments with plan characteristics information.	*Frictions:* Search costs to find or explore plan options. *Mental gaps:* People have limited insurance competence, not understanding the mapping between plan characteristics (for example, deductibles) and payoff-relevant outcomes.
Handel (2013) Abaluck and Gruber (2011, 2016); Ho, Hogan, and Scott Morton (2017); Ketcham, Lucarelli, and Powers (2015)	Consumer inertia leads to thousands of $ in financial losses (~$2,000) in insurance plan choice. Consumers choose dominated health plans with high frequency when possible to do so. Consumers leave money on the table in initial Medicare Part D choices, on average ~$300 per consumer. Consumers exhibit substantial inertia, leading to additional monetary losses.	*Frictions:* Switching costs (from search, information processing, etc.); rational inattention to plan choice. *Mental gaps:* Consumers don't recognize potential benefits from switching, having wrong priors about plan changes over time (for example, not realizing that plans may become financially dominated); lack of competency in evaluating premiums relative to plan characteristics; neglect of certain key plan features.
Health treatment		
Bronnenberg, Dubé, Gentzgow, and Shapiro (2015)	Experts (pharmacists and medical professionals) are less likely to pay extra for branded headache-remedy drugs relative to generic (bio-equivalent) alternatives.	*Frictions:* Information processing or search costs lead to unawareness of bio-equivalent alternatives. *Mental gaps:* People don't know which ingredients to focus on or realize that generic equivalents might be available; wrong priors about generic equivalence.
Pauly and Blavin (2008); Baicker, Mullainathan, and Schwartzstein (2015); Choudhry et al. (2011)	Health insurees seemingly underuse valuable treatments for chronic diseases. In such cases, health insurees' adherence is quite sensitive to copay changes.	*Frictions:* Information gathering and processing costs are too high for insurees to recognize the value of these treatments. *Mental gaps:* People do not know how to assess the value of treatments.

regimens aimed at preventing future heart attacks at regular copay levels, but show in a large-scale field experiment that eliminating copays for these drugs substantially boosts adherence and improves clinical outcomes. Baicker, Mullainathan, and Schwartzstein (2015) argue that it is difficult to rationalize such examples in a framework where consumers accurately trade off health benefits against the copay. Others argue that consumer misinformation is likely a key reason why consumers act as if they misweight treatment benefits (for example, Pauly and Blavin 2008). These findings have important policy implications: in many cases, when consumers make poor health choices it both increases long-run health costs and reduces consumer health, and everyone loses. How should policymakers use the evidence in these studies, or work to produce evidence in future studies, when considering different interventions to improve health care decisions?

Another set of examples comes from consumers' choices of health insurance plans. Handel (2013) analyzes this choice assuming that consumers have a bias toward inertia, modeled as costs from switching plans, but have rational expectations about their own health risk and full information about the plan options available. The paper estimates a switching cost of approximately $2,000 in the population. Many consumers leaving that much money on the table earn low incomes and have families, heightening the consequences. Handel acknowledges that the estimated switching cost likely reflects a range of underlying mechanisms, including true switching costs, search costs, and miscalibrated beliefs. Ho, Hogan, and Scott Morton (2017) model inertia using rational inattention as opposed to switching costs. They also find substantial inertia, modeled as a high cost of paying attention to the choice environment, with substantial negative consequences across the board for seniors. These two papers with similar data and identification assume distinct mechanisms underlying inertia, without teaching us which mechanism carries greater weight in the decision process.

Consumer Ignorance and Misinformation in Other Domains

Table 2 provides a few examples outside the health care arena. In one example, Hanna, Mullainathan, and Schwartzstein (2014) develop and test a model of technological learning. Focusing here on the empirical exercise, they study the knowledge, practices, and impact of a knowledge intervention on a community of Indonesian farmers who had a lot of experience: they farmed seaweed on average for 18 years with many cycles in each year. Seaweed is farmed by attaching strands of seaweed (or "pods") on lines submerged into the ocean, where many factors could affect yield. Local nongovernment organizations suggested that these farmers' practices tend to be far from the productivity frontier, a fact supported by Hanna et al.'s experimental estimates. Further, this appears to stem from farmers not understanding key relationships between input choices and yield. Farmers did precise things and had clear opinions on most dimensions: the length of their line, the distance between pods, the distance between lines, and the cycle length. But they did not have a clear opinion on their pod size (a truly important input dimension, according to Hanna et al.'s estimates): around 85 percent did not know the size they use and would not

Table 2
Examples of Information People Don't Use in Non-Health Markets

Paper	*Facts*	*Potential explanations for not using information*
Agriculture		
Hanna, Mullainathan, and Schwartzstein (2014)	Seaweed farmers persistently neglect an important input dimension (pod size). They respond to an intervention that filled in knowledge gaps.	*Frictions:* Learning potentially payoff-relevant relationships is costly. *Mental gaps:* Farmers started with wrong beliefs about which inputs mattered.
Financial investments		
Hortaçsu and Syverson (2004)	There is significant price dispersion in S&P 500 index funds (financially undifferentiated products). Higher-fee funds have meaningful market shares.	
Hastings, Hortaçsu, and Syverson (2017)	Consumers lose significant sums of money choosing among privatized, essentially homogeneous, mutual funds in Mexico's privatized social security. Advertising investment is associated with these poor choices.	*Frictions:* Large search costs to find prices or products; switching costs across firms. *Mental gaps:* People don't realize that index funds differ only in fees; advertising or marketing of brands may reinforce or cause wrong beliefs; limited financial literacy; people don't think to check on their 401(k) contribution rate.
Choi, Laibson, and Madrian (2010)	Consumers leave significant sums of money on the table by choosing high-fee index funds. Experiment shows this is not because of nonportfolio features and also is not primarily the result of search costs. Consumers with lower financial literacy are more likely to make mistakes, and often even have a sense they are making mistakes.	
Madrian and Shea (2001)	Consumers exhibit substantial inertia in their choice of 401(k) investments and are highly sensitive to default investment settings.	
Cellular phones		
Grubb and Osborne (2015)	Consumers demand cell phone plans as if they underestimate the variance of future calling minutes. Consumers appear inattentive to past usage within a plan month, making usage alerts valuable.	*Frictions:* Keeping track of usage is costly; switching costs in plan choice. *Mental gaps:* People underestimate the likelihood of using enough minutes to incur fees.
Energy		
Allcott and Taubinsky (2015)	Providing information on energy cost savings boosts demand for energy-efficient lightbulbs.	*Frictions:* Search costs for finding relevant product information. *Mental gaps:* People may be biased towards believing the upfront price is most important; people may focus too little on future costs.
Ito, Ida, and Tanaka (2016); Jessoe and Rapson (2014)	Consumers choose electricity tariffs that are bad for them as well as for society. Experiment shows that information provision helps reverse some of the poor decisions, but not a significant portion of them. High-frequency information provision makes consumers significantly better in responding to time-varying electricity tariffs and builds habits whereby consumers adjust behavior in the medium to long run even in the absence of information.	*Frictions:* Search costs of finding relevant electricity tariff information; switching costs of switching electricity plans; adjustment costs of changing electricity consumption in response to price fluctuations. *Mental gaps:* People may have low literacy in evaluating complex multipart electricity tariffs, or real-time electricity pricing; people may believe that information is hard to obtain when it is in fact easy to obtain.

give an opinion on the optimal size. This lack of opinion appeared to translate into a lack of measurement: Each farmer had substantial variation in pod size within his own plot (which in theory he could learn from). The failure to optimize pod size appeared to meaningfully reduce farmers' output and income.

In household finance, Hortaçsu and Syverson (2004) show that consumers frequently purchase higher-fee S&P 500 index funds as if they do not know of the existence of lower-fee funds that will provide essentially equivalent returns. The authors pose a model with consumer search frictions and assume that these search costs are responsible for the low-value options consumers end up choosing. Madrian and Shea (2001) study 401(k) decisions of many employees at a large firm and show that a shift in the default policy for how contributions are matched and invested has a substantial impact on consumers' investment strategies. Choi, Laibson, and Madrian (2010) dive into the mechanisms behind why individuals invest in index funds that do not minimize fees and show that this continues to hold when search costs are removed and is not explained by nonportfolio services. Hastings, Hortaçsu, and Syverson (2017) show that consumers in Mexico are heavily persuaded by advertising and pay substantial fees since the public pension system was privatized in the 1990s. These papers show broadly that consumers often leave a lot of money on the table in this domain, arguably because of misinformation, but still only scratch the surface of determining precisely why.

Table 2 highlights several other examples. Consumers act as if they do not know the features of certain options, such as the energy cost savings associated with energy-efficient lightbulbs (Allcott and Taubinsky 2015). They act as if they do not know add-on prices such as the sales taxes and shipping costs associated with consumer products (Chetty, Looney, and Kroft 2009; Brown, Hossain, and Morgan 2010). They act as if they do not know basic features of income tax schedules, such as marginal tax rates at current income levels (for example, Rees-Jones and Taubinsky 2016). They act as if they do not know information about their own behavior, like the number of cell-phone plan minutes they have used within a plan month (Grubb and Osborne 2015).

Discussion

While we have focused on a subset of markets, the evidence suggests that researchers would find that consumers face similar challenges in markets that have not yet been studied empirically, whether because of a lack of data or because it is difficult for researchers to assess mistakes in a given context. For example, it is simpler as a researcher to study branded versus generic drugs, which are chemically equivalent, than it is to study decisions where consumer heterogeneity is more important. But the finding that consumers overpurchase branded drugs suggests that consumers make misinformed choices in a variety of similar contexts. Likewise, the documented difficulties consumers have in choosing health insurance and financial products suggest that they also likely experience similar difficulties in choosing other complex financial products, such as life insurance, car insurance, credit cards, or loans.

As Tables 1 and 2 illustrate, "not knowing" in many of these examples could arise from a range of mechanisms. To think about how we might go about trying to distinguish between them (and the situations in which doing so is more or less important), it is useful to elaborate on what these mechanisms might be.

Possible Mechanisms

To help spell out possible mechanisms, consider the following framework. A person wishes to choose an action that maximizes utility. For example, the person could be choosing between health insurance plans, or between branded or generic drugs, or inputs to production that yield utility-relevant outcomes. This person faces uncertainty about the optimal action, such as uncertainty about prices, attributes of options, or the relationship between the action and outcome. However, the person can gather and process information that helps resolve this uncertainty. We'll simplify this discussion by collectively referring to the process of gathering and processing data as "attending to data."

In this setting, as one example, the person chooses a health insurance plan given attended-to information on prices and features of plans. Any strategy for attending to information includes a probabilistic distribution over information the person ultimately processes, and induces some potential cost to the person in terms of time and effort. The person should trade off the expected benefits of attending to information, b, against the costs of attending, c, thereby attending if $b > c$. In a number of settings, the benefits b of attending appear to be large, but the person doesn't seem to be attending to information. What could be going on?

The cost frictions framework says the costs of attending, c, must be large as well. For example, a consumer shopping in an insurance exchange may have correct beliefs about the distribution of prices in the market but incur cost (time and hassle) in finding and exploring each option in the choice set. Or the consumer may have all information on the insurance choice easily available but may not want to do the full calculation on expected costs given the nonlinear contract or health risk because it is too complex or time consuming. Models focusing on cost frictions in gathering, attending to, and integrating information include McCall (1970), Sims (2003), Gabaix (2014), and Woodford (2012).

But this isn't the only possibility. In the alternative mental gaps view, the person may misweight the benefits to attending, using some $\hat{b} \neq b$ in evaluating whether to attend, because important features of the problem are not at the top of the mind. For example, the consumer in a health insurance exchange may mistakenly believe the benefits from searching or attending to information about different options is low when in fact there is substantial price dispersion (or there have been substantial changes to the market). Alternatively, in considering employer plans, the consumer may believe that it is important to focus on the size of provider networks when instead the focus should be on deductibles and premiums. Similarly, a seaweed farmer may not appreciate that pod size matters much for yield. Models focusing on mental

gaps in gathering, attending to, and integrating information include Schwartzstein (2014) and Gagnon-Bartsch, Rabin, and Schwartzstein (2017). Recent laboratory experiments by Enke and Zimmerman (2017) and Enke (2017) explore mental gaps in some detail, as well as de-biasing strategies. Closely related for our purposes are models where a person overreacts to certain salient features of a problem, such as differences in deductibles. Models focusing on systematic errors in integrating information include Bordalo, Gennaioli, and Shleifer (2013), Kőszegi and Szeidl (2012), and Bushong, Rabin, and Schwartzstein (2017).

While not a focus of this article, there are several other possibilities for why people might act as if they are not attending to important information, even when $\hat{b} = b$ and c is low. First, it is of course possible that we as analysts are mismeasuring the potential benefits of improved attention to information. Second, the person may be motivated not to attend to information in order to preserve optimistic beliefs, for example about their own health status (Caplin and Leahy 2001; Brunnermeier and Parker 2005; Kőszegi 2006; Karlsson, Loewenstein, and Seppi 2009; Oster, Shoulson, and Dorsey 2013; Bénabou and Tirole 2016). Third, the person may act on the "wrong" decision utility function, placing too little weight on future benefits (Laibson 1997; O'Donoghue and Rabin 1999) or mispredicting future utility (for example, Loewenstein, O'Donoghue, and Rabin 2003).

Table 3 presents a more detailed look at the frictions and mental gaps mechanisms together with examples from the literature, more carefully decomposing the choice process into stages when barriers to acquiring and optimally using information arise. Some of the examples discussed earlier arguably reflect either mostly frictions or mostly mental gaps. But turning back to Tables 1 and 2, the final column illustrates how many of the examples discussed earlier are consistent with both. Consider the Bronnenberg et al. (2015) branded versus generic drugs example. Cost frictions could be at play: it may be costly to find the generic alternative on the store shelf or to learn about active ingredients. Mental gaps may also play a role: people may not appreciate that generic alternatives to headache remedies are available or believe that chemical equivalence between the products is a possibility worth exploring. Distinguishing between mechanisms in examples such as these requires a more nuanced approach.

Empirical Approaches to the Magnitude of and Reasons for Error

This section discusses empirical approaches for studying environments where cost frictions and mental gaps are present. In this discussion, we will assume that we are considering situations where such frictions and gaps are the primary drivers of the wedge between choices people "should" make and choices they in fact make.

Total Impact on Demand

A range of empirical work seeks to identify the demand curve that represents consumers' actual choices separately from the demand curve in a frictionless

Table 3
Some "Whys" of Not Using Information

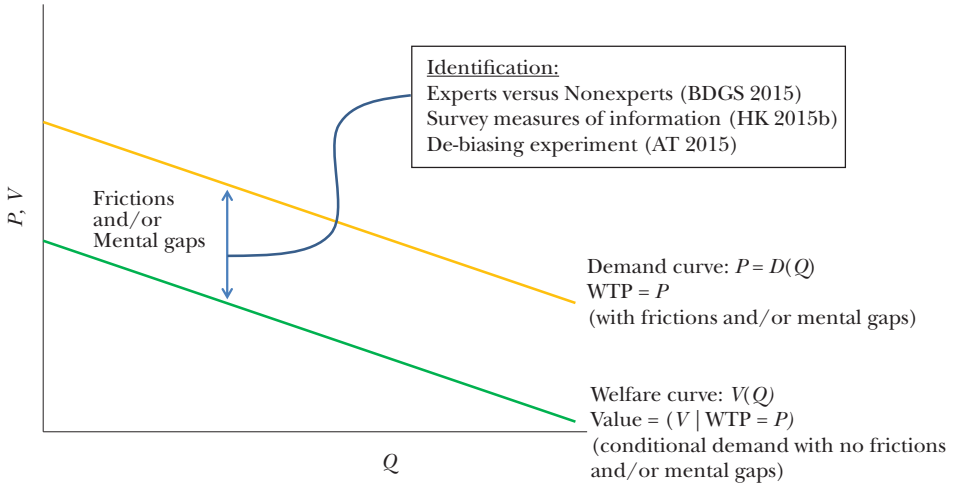
Frictions:		
When gathering	*When attending*	*When integrating*
Search costs + Rational expectations *Models:* Stigler (1961); McCall (1970); Caplin and Dean (2015) *Examples:* Hortaçsu and Syverson (2004)—Mutual funds; Cebul, Rebitzer, Taylor, and Votruba (2011)—Health insurance; Ellison and Ellison (2009)—Online markets	**Rational inattention:** *Models:* Sims (2003); Gabaix (2014); Matějka and McKay (2015) *Examples:* Bartoš, Bauer, Chytilová, and Matějka (2016)—Labor market discrimination	**Costs of complex thinking, difficulty doing math** *Models:* Ortoleva (2013) *Examples:* Handel and Kolstad (2015b)—Health insurance choice
Mental gaps:		
When gathering	*When attending*	*When integrating*
Search with subjective priors *Models:* Rothschild (1974); Rosenfield and Shapiro (1981) *Examples:* De los Santos, Hortaçsu, and Wildenbeest (2012)—Web browsing and purchasing	**Noticing / Selective attention** *Models:* Schwartzstein (2014); Gagnon-Bartsch, Rabin, and Schwartzstein (2017) *Examples:* Hanna, Mullainathan, and Schwartzstein (2014)—Farming; Malmendier and Lee (2011)—eBay bidding	**Salience, focusing, relative thinking, limited financial literacy** *Models:* Bordalo, Gennaioli, and Shleifer (2012, 2013); Kőszegi and Szeidl (2012)—Salience and focusing; Bushong, Rabin, and Schwartzstein (2017)—Relative thinking *Examples:* Bhargava, Loewenstein, and Sydnor (2017)—Health insurance choice

environment with fully rational consumers, which is called the "welfare-relevant" curve. Understanding and estimating the wedge between these two demand curves is sufficient for a variety of important policy questions (Mullainathan, Schwartzstein, and Congdon 2012). Again, we will equate "welfare" with consumer welfare throughout our discussion.

Figure 1 illustrates the demand curve and the welfare-relevant valuation curve for a hypothetical product. The welfare curve is defined conditional on the demand curve, such that the value of the welfare curve shown at any point reflects the average value for marginal consumers on the demand curve at a given quantity level.[2] The wedge between them represents the case where demand is higher than in a rational frictionless environment, leading to over-purchasing in an allocative sense. Each of

[2]For simplicity, we will refer to the "demand curve" and "welfare curve" as the key sufficient objects. For certain policy cases, discussed in more depth in the next section, the researcher will also want to understand heterogeneity conditional on a given level of demand in order to use these objects to evaluate policies where consumers may have heterogeneous responses—for example, to taxes or subsidies that may not be equally salient for everyone.

Figure 1
Demand versus Welfare-Relevant Valuation



Notes: This figure illustrates empirical approaches that seek to identify observed demand, including frictions and mental gaps, from the welfare-relevant valuation curve, which in some contexts is equivalent to the demand curve for fully informed, frictionless, and bias-free consumers. The welfare-relevant valuation curve gives what true experienced product values would be for consumers at a given level of demand. The wedge between the demand and welfare curve can be due to a range of underlying mechanisms (for example frictions and/or mental gaps) and there are several identification approaches used in the literature to identify this gap. BDGS stands for Bronnenberg et al., HK for Handel and Kolstad, and AT for Allcott and Taubinsky. WTP is "willingness to pay."

the frictions or mental gaps described in the previous section could contribute to this wedge. Recent empirical research highlights several different options for identifying both the demand and welfare-relevant valuation curve in a given environment.

A first empirical strategy estimates a demand curve for experts and a separate demand curve for nonexperts based on the assumption that the demand curve for experts represents the demand curve in a rational frictionless world for experts *and* nonexperts, conditional on a range of observables. In a study mentioned earlier, Bronnenberg et al. (2015) take this approach in studying demand for generic drugs relative to their branded counterparts. When they have quantified the wedge between true demand (of nonexperts) and the welfare-relevant valuation (of experts) for branded versus generic drugs, they can then use this calculation as an input into a welfare analysis of various policies that shift consumers towards generic drugs.

A second approach to identifying this wedge, based on a similar intuition, uses a survey that separates informed from uninformed consumers. The underlying assumption is that informed consumers as measured by the survey make rational full information choices in the context of a neoclassical expected utility model.

One can then quantify the wedge between the demand for informed and uninformed consumers. Handel and Kolstad (2015b) take this approach in seeking to understand why consumers under-purchase high-deductible health plans in a large-employer health insurance context.

A third approach involves using a randomized trial to create a class of well-informed consumers, who can then be compared to others. Allcott and Taubinsky (2015) take this approach in studying the demand for energy-efficient lightbulbs, which seem to be under-purchased relative to both their value for a given individual and relative to their social value (given the externalities imposed by inefficient energy consumption). They assume that consumers in the treatment group are "fully de-biased"—that is, equivalent to the rational frictionless experts and fully informed consumers in the previous two methods. Under this assumption, the demand curve for treated consumers represents the welfare-relevant value curve for all consumers conditional on key observable factors, while actual demand including mental gaps and frictions can be estimated using the control group.

These three approaches differ in the assumptions required to identify welfare-relevant valuation separately from demand.[3] The first strategy (comparing acknowledged experts to nonexperts) is probably the most robust approach of the three, assuming that experts can be appropriately differentiated. Here, the assumptions are that for experts the cost of attention c is relatively low and the perceived benefits to attention are similar to the actual benefits, $\hat{b} \approx b$.

The second approach (using a survey to identify informed and uninformed consumers) presumes that informed consumers have similar preferences to uninformed consumers (conditional on detailed observables), but because they are better informed, they are able to accurately link those preferences to choices. One weakness of this approach relative to the first approach is that eliciting preferences and information sets via survey can introduce well-known issues of measurement error (for discussion, see Bertrand and Mullainathan 2001). A weakness of both approaches is that experts (or informed consumers) who look similar to nonexperts (or uninformed consumers) on observable characteristics may be different on unobservable characteristics.

The third, "de-biasing experiment," approach assumes that the treatment gives a consumer the expertise necessary to operate as a rational frictionless agent (through better calibrating their estimates of benefits $\hat{b}$ or by reducing costs c). The assumptions in this approach are likely the strongest of those needed across the three approaches; indeed, in some cases, the "de-biasing" may even overshoot the true demand curve for reasons argued by Bordalo, Gennaioli, and Shleifer (2015). This approach assumes that the intervention *causes* expertise in a domain, rather than measuring it (survey) or verifying it (occupation data). Of course, experiments can be combined with detailed surveys to assess the level of information or biases

[3] A fourth approach, explored by Baicker, Mullainathan, and Schwartzstein (2015), is to estimate (or bound) the welfare curve by directly measuring proxies for inputs to welfare, such as health outcomes.

a consumer has, potentially improving on approaches that use one method or the other.

All three of these approaches assume that a constellation of cost frictions and mental gaps drive the wedge between choices people "should" make and choices they do make. Along with biases specifically related to information, other biases may also be at play in some of the decisions studied, such as present-bias (Laibson 1997; O'Donoghue and Rabin 1999). Greater knowledge of the mechanism(s) driving the wedge in a particular application can bolster confidence in estimates of its size.

Empirical Identification of Specific Mechanisms

The majority of papers that seek to estimate a wedge between demand and welfare curves suggest a specific mechanism that may have caused the wedge, but rarely test their suggested explanation against other possible explanations. For example, a paper that estimates search, switching, or attentional costs typically models a consumer with beliefs closely tied to a rational beliefs framework who incurs costs to acquire key information and improve choices (for example, Hortaçsu and Syverson 2004; Handel 2013). While such papers acknowledge that other factors could also drive poorly informed choices, typically these other factors are not explicitly included in the model. A range of other papers, which are typically less-structural, alternatively allow consumers to make mistakes but largely abstract from more traditional search or processing costs that a social planner might not want consumers to incur (for example, Baicker, Mullainathan, and Schwartzstein 2015).

Distinguishing between the potential mechanisms can provide a more precise characterization of demand versus welfare-relevant value in environments and enable more accurate predictions of policy impacts, but may also require additional data or empirical assumptions. Researchers have used several approaches to differentiate empirically between competing mechanisms. The first uses theoretically motivated assumptions in the context of structural models to test hypotheses about underlying mechanisms: for example, Malmendier and Lee (2011) use this approach to study why some consumers pay more for an item in an eBay auction than they would in a simultaneous fixed-price offering. They test for whether a combination of price uncertainty and switching costs can explain these patterns and argue that the data are inconsistent with this mechanism, implying that some additional mental gap must be a partial cause of these mistakes.[4] One feature of these and other structural approaches (for example, Grubb and Osborne 2015; Barseghyan, Molinari, O'Donoghue, and Teitelbaum 2013) is that they can distinguish between a specific set of potential mental gaps or frictions but must maintain assumptions about other gaps and frictions to do so. Ultimately, the credibility of

[4]Schneider (2016) comments on the Malmendier and Lee (2011) paper, suggesting that, under some assumptions, adding traditional search costs into the model can rationalize what might otherwise appear to be bidder mistakes on eBay.

each approach rests on how reasonable these assumptions are in the context of the specific environment being studied.

A second approach used more informally in the literature is to choose one specific mechanism to represent the set of frictions and biases, but then to use calibration arguments to argue that this mechanism is unlikely to explain the entire wedge between demand and welfare-relevant value. For example, Handel's (2013) model of health plan choice assumes that inertia—in which consumers stay with their previous plan even after the elements of the plan have shifted—might result from consumer switching costs. But the size of the switching costs needed to produce this result are estimated to be approximately $2,000. Based on typical values of time costs and some intuition about consumer valuation, this cost seems "too large." The author uses this observation to discuss other potential explanations for inertia in switching between insurance plans, such as biased beliefs and inattention.

A third option is to use survey data to elicit responses about different frictions or mental gaps. For example, Handel and Kolstad (2015b) ask questions about information on a range of dimensions to assess the contribution of different kinds of limited information to demand for health plans. They show, for example, that a lack of information about provider networks has a large impact on demand for high-deductible plans. Their primary structural framework includes indicators for limited information in a reduced-form way, and an alternative framework (presented in an appendix) structurally links indicators of limited information to biased beliefs about certain plan dimensions. Hanna, Mullainathan, and Schwartzstein (2014) similarly combine survey and behavioral data to differentiate between some reasons why seaweed farmers' practices are seemingly off the production possibilities frontier. While classical explanations would likely involve frictions to information-gathering—for example, perhaps due to costs of experimentation—the data instead suggest that farmers were not paying attention to key input dimensions in their own activities. As discussed above, a vast majority could not answer questions about their own practices with regard to key inputs.

A fourth option is to use "mechanism experiments" (Ludwig, Kling, and Mullainathan 2011) to understand the relative impacts of different frictions or biases. Bhargava, Loewenstein, and Sydnor (2017) explore the quality of individuals' health insurance decisions, and reasons for what appear to be mistakes, by analyzing data from an employer where employees choose from large menus of insurance plan options. They find that a *majority* of employees choose health insurance plans that are financially dominated: For example, an employee might pay $500 more in annual premiums to reduce the deductible from $1,000 to $750. One natural hypothesis is that consumers choose financially dominated options because search is difficult and consumers do not know that financially dominating options are available. But evidence from follow-up experiments suggests a basic error may be even more important: many consumers do not appear to know how to map insurance plan features into final wealth outcomes. In a follow-up survey done using the Qualtrics online survey platform, 66 percent of participants choose a financially dominated plan even when the presentation of options was highly simplified to

include four options that only varied in deductible and premium. On the other hand, in another follow-up experiment, this time on Amazon Mechanical Turk, clarifying the relationship between various premium and deductible combinations and total health costs reduced the fraction of participants choosing dominated plans from 48 to 18 percent. Further, those with higher measured understanding of health insurance concepts in this experiment were less likely to choose dominated plans.

When Do We Care Why?

In contexts where consumers appear to leave a lot of money on the table, an obvious accompaniment to looking at the welfare losses is to consider counterfactual public policies, which by definition are out-of-sample. For example, in the health insurance exchanges set up under the Patient Protection and Affordable Care Act of 2010, it is very costly to change regulations related to consumer choice environments (for example, specifying a set of allowable contracts, web designs, or ways in which benefits are represented) and useful to predict impacts of potential new policies.

As you recall, *allocation policies* directly allocate (or strongly steer) consumers to specific actions, and so the underlying cause of the error is unlikely to matter much for policy analysis. *Mechanism policies* instead target specific mechanisms, and so the underlying cause of the consumer error will matter for analysis of that type of policy. While these definitions are not intended to be mutually exclusive or exhaustive, they are intended to broadly frame policies as those that either do or do not strongly interact with the mechanism underlying poorly informed choices.

Allocation Policies

Regulations that remove specific poor options from choice sets, force or nudge consumers into certain better products, or use targeted default options are all examples of allocation policies. Traditional price instruments, such as taxes and subsidies (assuming consumer awareness of those taxes and subsidies) can also be allocation policies. For allocation policies, knowing the precise mechanism behind poorly informed choices is arguably less important than knowing the existence and magnitude of the consumer error.

Table 4 provides examples of some allocation policies. One example in health insurance markets is plan regulation that restricts the actuarial value of plans that insurers can offer in the market. Exchanges set up under the Patient Protection and Affordable Care Act of 2010 allow insurers to offer plans in four tiers of actuarial value: 60, 70, 80, and 90 percent of expected healthcare costs. Consider potential policies that either 1) raise the minimum allowable coverage to 70 percent actuarial value or 2) reduce the maximum allowable coverage to 80 percent actuarial value. Though there are some potential equilibrium pricing consequences that result from such regulation, the first-order effect is likely to shift an entire population of

Table 4
Allocation versus Mechanism Policies

Allocation Policies	*Mechanism Policies*
Health insurance	
Market regulation in Affordable Care Act regarding plan design, like minimum cost-sharing, or structure of cost-sharing. Regulation of minimum networks and covered services. Changes to default insurance options or processes (for example, targeted defaults).	Choice-framing in insurance markets through web design and information display. Education campaigns to promote insurance literacy. Availability of aggregate and disaggregate information on insurer networks. Standardized representation of insurance plans.
Health care services	
Mandatory generic substitution for drugs. Changing medical guidelines to induce changes in default treatments for patients. Value-based cost-sharing.	Information pamphlets and posting about equivalence of brand and generic drugs. Choice framing for brand versus generic drugs. Shared decision making for difficult medical decisions. Information provision on costs or outcomes of medical services.
Financial investment	
Fee regulation eliminating plans with certain types of hidden fees. Default options in 401(k) choices.	Education campaigns promoting financial literacy. Standardized display of key fund features. Improvements to search tools.
Energy-efficient products	
Regulation on level of energy efficiency required for products. Taxing energy-inefficient products or subsidizing efficient products.	Education about the value energy efficiency can provide financially. Education about the impacts of energy efficiency on the environment.
Agricultural production	
Subsidizing or directly distributing inputs like fertilizer.	Agricultural extension, outreach, and education.
School choice	
Changes to the default options or the steering inherent in the choice mechanism. Limiting the set of available schools.	Information provision about school-choice mechanism, or school options. Changes to the complexity of the mechanism.

consumers either up or down in terms of coverage generosity. The welfare implications could be assessed if the demand curve and welfare curve for one coverage tier relative to another are identified, without appealing to the specific mechanisms driving the wedge between these curves. Handel and Kolstad (2015b) study a similar example where a large employer shifts from offering multiple insurance options to just one option, a high-deductible plan. The authors are able to assess the welfare implications of such a move after identifying the relevant demand and

welfare curves prior to this forced switch. Because most employers offer only one insurance plan, and many switch their plans over time, forced choice is especially relevant in that market.

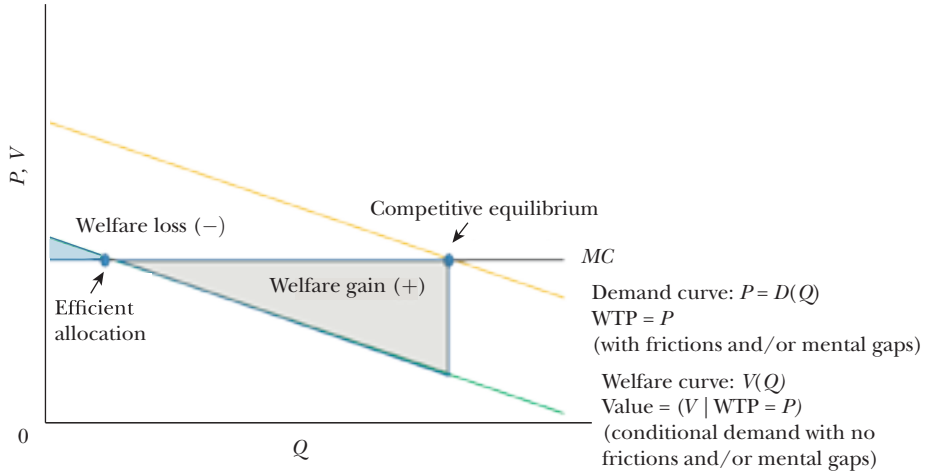
In health care services more broadly, a number of states have implemented mandatory generic substitution laws, which essentially require mandatory substitution from brand drugs to generic equivalents except in certain exceptional circumstances. Work like Bronnenberg et al. (2015) that identifies the demand curve and welfare curve for purchases of brand versus equivalent generic drugs can help predict the welfare effects of such a policy. In the domain of over-the-counter drugs, where consumers may have more discretion than for prescribed drugs, the estimates of Bronnenberg et al. also inform how we might want to tax branded drugs or subsidize generic drugs. In health treatment markets, Baicker, Mullainathan, and Schwartzstein (2015) argue that knowing the extent to which people on the price margin are underusing certain treatments (for example, drugs to prevent future heart attacks) is enough to conclude that it would be welfare-enhancing to reduce prices, even without knowing exactly what leads to such underuse.

Table 4 lists examples of allocation policies related to financial investments, energy-efficient products, agricultural production, and school choice. Across these sectors, and the others already discussed, we include default policies that strongly influence the allocation of consumers to products as a borderline case of allocation policies. For example, Madrian and Shea (2001) and follow-up work illustrate how changing the default choice of whether one is automatically enrolled in a retirement savings program powerfully affects the extent of consumer savings. While the effect of defaults of course depends somewhat on the mechanism that drives a wedge between the demand curve and the welfare curve, arguably it is broadly independent of precise details of this mechanism. Table 4 includes a number of other contexts where default policies are likely to be close in spirit to allocation policies.

Figure 2 illustrates the welfare implications of an allocation policy in the context of a market for a commodity. For example, assume that the purchase decision studied is the case where consumers are considering whether to buy a brand drug or a generic drug, but that consumers on average are biased towards purchasing a brand drug, relative to actual benefits. The demand curve represents the *relative* revealed preference for a brand drug relative to a generic drug, as a function of price, while the welfare curve represents the distribution of the *actual welfare-relevant relative value* for fully informed, frictionless, and bias-free consumers. The cost curve represents the higher social marginal cost of the brand drug, perhaps in this case because of advertising.

The figure illustrates the welfare effects of an allocation policy that allocates all consumers to the generic counterpart of a branded drug. Consumers who had been purchasing branded drugs, but for whom the actual relative value of the branded drug is much lower, have a large welfare gain. But the figure also allows for the possibility that some subset of consumers loses from this allocation policy: even if all consumers have the same bias towards purchasing branded drugs, as the figure posits, some subset might still value the branded drug above its relative marginal cost.

Figure 2
Welfare Impact of an Allocation Policy
(for instance, forcing consumers to buy a generic drug rather than a brand drug)



Notes: This figure illustrates the welfare impact of an allocation policy that restricts the quantity consumed to zero in a market where there is a wedge between demand and welfare-relevant valuation, as a result of frictions and/or mental gaps. The figure applies to the simple case of a competitive market for two products with constant marginal costs, for example, as in the Bronnenberg et al. (2015) case of consumers who consider whether to purchase a brand drug or a generic equivalent. In that case, the demand and welfare curves reflect the relative willingness-to-pay and valuation for a brand drug compared to its generic counterpart, and quantity reflects the amount of the branded drug consumed.

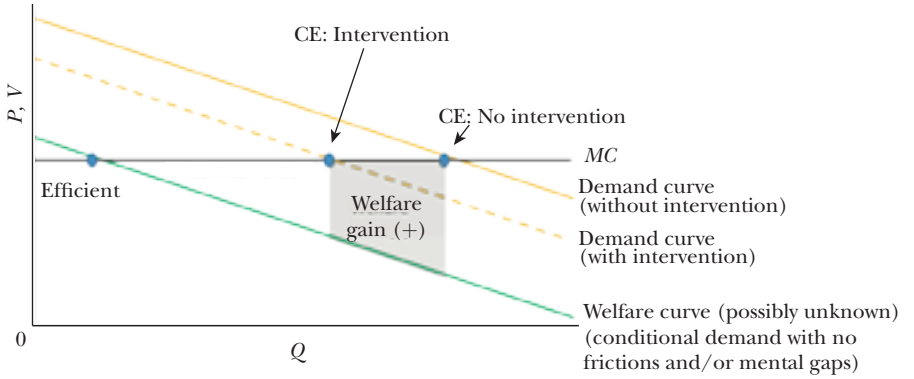
This could, for example, be a case where consumers have a placebo effect induced by advertising for a branded drug, or just gain utility from purchasing a heavily advertised product. This figure underscores that in order to predict the welfare effect of an allocation policy, assessing heterogeneity in perceived value (demand), actual value, and the overall extent of frictions or mental gaps are crucial, while differentiating between specific frictions and mental gaps may be less important.

The distinctions in Figure 2 should be viewed as approximations. In some empirical contexts the data and identification strategy for determining the wedge between demand and welfare-relevant demand do not allow for tight estimates. Relatedly, the definition of an allocation policy is one of degree. Clearly, mandatory generic substitution moves consumers to generic drugs in a way that leaves little room for consumer mental gaps or frictions to affect the outcome of the policy. However, the identified demand and value curves may not be sufficient for studying a tax or subsidy policy if, for example, taxes and subsidies are not particularly "salient" for some consumers (Chetty, Looney, and Kroft 2009). To evaluate such policies, the researcher needs to analyze heterogeneous consumer responses, which may be a function of underlying mechanisms (Taubinsky and Rees-Jones forthcoming).

Figure 3

Welfare Impact of a Mechanism Policy

(for instance, providing consumers with information about the relative value of branded drugs and generic drugs)



Notes: This figure shows the impact of a mechanism policy, for example one that provides information to consumers about the relative value of branded drugs compared to generic drugs. It shows the case where the policy has a homogeneous impact on all consumers, shifting them part-way towards the true welfare curve from the demand curve. CE stands for competitive equilibrium.

Mechanism Policies

Mechanism policies target specific frictions or mental gaps. For example, sending a consumer a targeted message with choice-relevant information may effectively promote better outcomes if information availability or search costs were the first-order problem, but will be ineffective if the information were always readily available and mental gaps having to do with using or processing that information are more material.

Table 4 also lists some examples of mechanism policies. In health insurance markets, for example, these include standardized representation of insurance plans (Ericson and Starc 2016), education campaigns to promote insurance literacy, choice-framing via specific choice orderings and web designs, or intensive targeted information provision (Kling, Mullainathan, Shafir, Vermeulen, and Wrobel 2012; Handel and Kolstad 2015a). Table 4 also lists some policies related to energy, school choice, and agricultural production. In order to predict the effects of policies that target a specific information-related friction or mental gap, it is necessary to identify the role that mechanism plays in driving choices and potential mistakes. As discussed earlier, this can be quite difficult, usually requiring either multi-arm experiments, comprehensive linked surveys that target information acquisition and processing issues, or natural experiments linked with structural assumptions about these microfoundations.

Figure 3 illustrates the welfare impact of a mechanism policy. For simplicity, the figure assumes that the mechanism policy impacts all consumers evenly, though this

is unlikely to be the case in reality. One can imagine this example as representing the case of information provision for the relative quality of branded drugs versus generics, which would likely reduce the relative demand of some consumers for branded drugs. The figure illustrates the demand impact of this policy, assuming that limited information is one reason, but not the entire explanation, for the wedge between demand and welfare-relevant valuation.

The figure shows the potential pitfalls of using a mechanism policy, like a helicopter drop of information, without having a good sense of the mechanism beforehand. First, the policy may not be very effective: in this case, if information frictions are but one of several frictions and mental gaps, then the drop in demand from the policy is small relative to the drop if the policy truly eliminated all frictions and mental gaps. Second, if the policy used to *remove* frictions and fill in mental gaps was also used in earlier research to *measure* the magnitude of frictions and mental gaps, then the results could seriously understate the benefits from trying hard to eliminate all frictions and gaps. Third, if the demand curve under the policy is mistakenly viewed as the welfare curve, then not only will we understate the potential welfare gains from an ideal policy, we will understate the welfare gains from *this* policy. A given fall in demand from an information drop may appear to barely raise welfare not only because the fall is small, but because this small effect could mislead researchers to infer that people were making good choices to begin with.

In many cases—such as with providing information, making an interface simpler, or encouraging the consumer to make an active choice—it is useful to remember that even policies that seem blunt or obvious may not necessarily target the relevant mechanisms.

An additional key issue in mechanism policies is the extent to which changing the nature of consumer engagement with the choice process causes them to incur additional costs. For example, a policy that reduces consumer information processing costs—for example, via standardized presentation of product attributes—may have multiple effects: 1) help consumers make better choices; 2) cause them to devote more time to the choice process; and 3) incur *more* processing costs than before as a result of this increased engagement. A more straightforward example is a policy that encourages active choices (Carroll, Choi, Laibson, Madrian, and Metrick 2009). For such policies, it is important not only to understand how those policies might improve outcomes (as shown in Figure 3) but also to understand how the costs incurred during the choice process change (and Figure 3 abstracts from that change).

The above policy discussion comes from the perspective of an analyst who seeks to evaluate the likely impacts of a counterfactual policy, whether that policy is an allocation or mechanism policy. It is also possible to evaluate the welfare impacts of a mechanism policy without identifying the exact underlying mechanisms in the case where the empirical analyst can both separate true consumer value from willingness-to-pay and also evaluate the positive impacts of a mechanism policy on choices. In this case, the researcher can use the techniques described (like comparisons of

experts versus nonexperts) to identify value from demand, and can use these fundamentals together with the empirical implementation of a policy to assess its welfare impacts. This can be an efficient way to evaluate mechanism policies when testing these policies is simple and cheap, and when there is a clear way to identify true value from willingness-to-pay in the empirical environment.

Discussion

Rapid improvements in the depth and scope of data available to empirical research have fueled a wave of recent research on the extent to which consumers leave meaningful value on the table as a result of frictions and mental gaps. Policymakers have used this research to motivate a wide range of policies, including setting default options, influencing or constraining choice sets, providing information, standardizing products, and promoting active choices.

Yet there is much weaker evidence on which mechanisms are most important in given contexts. Many research articles explicitly model one mechanism as *the* key friction or mental gap and assume away all other potential explanations. This is typically done for simplicity and for exposition: it is often useful for researchers to act as if one mechanism were the true mechanism even if there is little in the data to distinguish it from other potential mechanisms. Such articles often discuss other potential mechanisms as alternative explanations but don't consider them in depth.

Our main goal in this article is to highlight this issue and investigate how to deal with it in empirical work and policy analysis. Economists sometimes have an intuition that nudges are more conservative than traditional policy instruments when we are uncertain about the mechanism underlying poor choices (Thaler and Sunstein 2009). In contrast, we emphasize a way that blunter allocation policies may actually be conservative: for allocation policies, it is less important to understand the precise mechanisms leading to consumer mistakes than to estimate the wedge between demand and welfare. A growing literature uses survey data, data on experts, and "de-biasing" experiments to identify this wedge and to illustrate its implications for different policies.

The ability to characterize the impacts of allocation policies more easily means that policymakers may have a more precise assessment of those policies, not necessarily that those policies are preferable. The direct intervention of allocation policies is a blunt instrument that may ignore heterogeneity in consumer preferences and the valuable role that informed consumers play in causing the market to provide the best possible products at the lowest possible prices.

More targeted mechanism policies may be better or more politically palatable. However, to evaluate the potential effects of these policies, it is crucial to understand which specific mechanisms lead to consumer mistakes in the first place. While noting the paucity of such evidence across important contexts, we highlighted some promising approaches. As data depth and scope improve, empirically disentangling mechanisms in a given context will become increasingly viable.

■ *For helpful comments, we thank Stefano DellaVigna, Benjamin Enke, Matthew Gentzkow, Brian Hall, Michael Luca, Ulrike Malmendier, Ann Norman, Matthew Rabin, Jesse Shapiro, Andrei Shleifer, Dmitry Taubinsky, Timothy Taylor, and Neil Thakral.*

References

Abaluck, Jason, and Jon Gruber. 2011. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program." *American Economic Review* 101(4): 1180–1210.

Abaluck, Jason, and Jon Gruber. 2016. "Evolving Choice Inconsistencies in Choice of Prescription Drug Insurance." *American Economic Review* 106(3): 2145–84.

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105(8): 2501–38.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein. 2015. "Behavioral Hazard in Health Insurance." *Quarterly Journal of Economics* 130(4): 1623–67.

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C Teitelbaum. 2013. "The Nature of Risk Preferences: Evidence from Insurance Choices." *American Economic Review* 103(6): 2499–2529.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka. 2016. "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition." *American Economic Review* 106(6): 1437–75.

Beaman, Lori, Jeremy Magruder, and Jonathan Robinson. 2014. "Minding Small Change among Small Firms in Kenya." *Journal of Development Economics* 108: 69–86.

Bénabou, Roland, and Jean Tirole. 2016. "Mindful Economics: The Production, Consumption, and Value of Beliefs." *Journal of Economic Perspectives* 30(3): 141–64.

Bertrand, Marianne, and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *American Economic Review* 91(2): 67–72.

Bhargava, Saurabh, George Loewenstein, and Justin R. Sydnor. 2017. "Choose to Lose: Health Plan Choices from a Menu with Dominated Option." *Quarterly Journal of Economics* 132(3): 1319–72.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice under Risk." *Quarterly Journal of Economics* 127(3): 1243–85.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2013. "Salience and Consumer Choice." *Journal of Political Economy* 121(5): 803–43.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2015. "Memory, Attention and Choice." https://scholar.harvard.edu/files/shleifer/files/evokedsets_march18_0.pdf.

Bronnenberg, Bart J., Jean-Pierre Dubé, Matthew Gentzkow, and Jesse M. Shapiro. 2015. "Do Pharmacists Buy Bayer? Informed Shoppers and the Brand Premium." *Quarterly Journal of Economics* 130(4): 1669–1726.

Brown, Jennifer, Tanjim Hossain, and John Morgan. 2010. "Shrouded Attributes and Information Suppression: Evidence from the Field." *Quarterly Journal of Economics* 125(2): 859–76.

Brunnermeier, Markus K., and Jonathan A. Parker. 2005. "Optimal Expectations." *American Economic Review* 95(4): 1092–1118.

Bushong, Benjamin, Matthew Rabin, and Joshua Schwartzstein. 2017. "A Model of Relative Thinking." http://www.hbs.edu/faculty/Pages/download.aspx?name=RelativeThinking.pdf.

Caplin, Andrew, and Mark Dean. 2015. "Revealed Preference, Rational Inattention, and Costly Information Acquisition." *American Economic Review* 105(7): 2183–2203.

Caplin, Andrew, and John Leahy. 2001. "Psychological Expected Utility Theory and Anticipatory Feelings." *Quarterly Journal of Economics* 116(1): 55–79.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *Quarterly Journal of Economics* 124(4): 1639–74.

Cebul, Randall D., James B. Rebitzer, Lowell J. Taylor, and Mark E. Votruba. 2011. "Unhealthy

Insurance Markets: Search Frictions and the Cost and Quality of Health Insurance." *American Economic Review* 101(5): 1842–71.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99(4): 1145–77.

Choi, James, David Laibson, and Brigitte C. Madrian. 2010. "Why Does the Law of One Price Fail? An Experiment on Index Mutual Funds." *Review of Financial Studies* 23(4): 1405–32.

Choudhry, Niteesh K., Jerry Avorn, Robert J. Glynn, Elliott M. Antman, Sebastian Schneeweiss, Michele Toscano, Lonny Reisman, Joaquim Fernandes, Claire Spettell, Joy L. Lee, Raisa Leven, Troyen Brennan, and William H. Shrank for the Post-Myocardial Infarction Free Rx Event and Economic Evaluation (MI FREEE) Trial. 2011."Full Coverage for Preventive Medications after Myocardial Infarction." *New England Journal of Medicine*. Special article posted December 1. http://www.nejm.org/doi/full/10.1056/nejmsa1107913#t=article.

Cunningham, Tom. 2013. "Comparisons and Choice." Unpublished paper. May 29.

De los Santos, Babur, Ali Hortaçsu, and Matthijs R. Wildenbeest. 2012. "Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior." *American Economic Review* 102(6): 2955–80.

Ellison, Glenn, and Sara Fisher Ellison. 2009. "Search, Obfuscation, and Price Elasticities on the Internet." *Econometrica* 77(2): 427–52.

Enke, Benjamin. 2017. "Complexity, Mental Frames, and Neglect." Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2922274.

Enke, Benjamin, and Florian Zimmermann. 2017. "Correlation Neglect in Belief Formation." Unpublished paper, August 8.

Ericson, Keith M. Marzilli, and Amanda Starc. 2016. "How Product Standardization Affects Choice: Evidence from the Massachusetts Health Insurance Exchange." *Journal of Health Economics* 50: 71–85.

Farhi, Emmanuel, and Xavier Gabaix. 2017. "Optimal Taxation with Behavioral Agents." NBER Working Paper 21524.

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 129(4): 1661–1710.

Gagnon-Bartsch, Tristan, Matthew Rabin, and Joshua Schwartzstein. 2017. "Channeled Attention and Stable Errors." Preliminary, December 10. https://scholar.harvard.edu/files/gagnonbartsch/files/channeledattentioncleandec10.pdf.

Grubb, Michael D., and Matthew Osborne. 2015. "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock." *American Economic Review* 105(1): 234–71.

Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *American Economic Review* 103(7): 2643–82.

Handel, Ben, and Jonathan Kolstad. 2015a. "Getting the Most From Marketplaces: Smart Policies on Health Insurance Choice." Brookings Hamilton Project Discussion Paper 2015-8. October. http://www.hamiltonproject.org/assets/files/smart_policies_on_health_insurance_choice_final_proposal.pdf.

Handel, Benjamin R., and Jonathan T. Kolstad. 2015b. "Health Insurance for 'Humans': Information Frictions, Plan Choice, and Consumer Welfare." *American Economic Review* 105(8): 2449–2500.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2014. "Learning through Noticing: Theory and Evidence from a Field Experiment." *Quarterly Journal of Economics* 129(3): 1311–53.

Hastings, Justine S., Ali Hortaçsu, and Chad Syverson. 2017. "Sales Force and Competition in Financial Product Markets: The Case of Mexico's Social Security Privatization." *Econometrica* 85(6): 1723–61.

Ho, Kate, Joseph Hogan, and Fiona Scott Morton. 2017. "The Impact of Consumer Inattention on Insurer Pricing in the Medicare Part D Program." NBER Working Paper 21028.

Hortaçsu, Ali, and Chad Syverson. 2004. "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds." *Quarterly Journal of Economics* 119(2): 403–56.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2016. "Information Frictions, Inertia, and Selection on Elasticity: A Field Experiment on Electricity Tariff Choice." July 24. https://cenrep.ncsu.edu/cenrep/wp-content/uploads/2016/12/K-Ito.pdf.

Jessoe, Katrina, and David Rapson. 2014. "Knowledge is (Less) Power: Evidence from Residential Energy Use." *American Economic Review* 104(4): 1417–38.

Karlsson, Niklas, George Loewenstein, and Duane Seppi. 2009. "The Ostrich Effect: Selective Attention to Information." *Journal of Risk and Uncertainty* 38(2): 95–115.

Ketcham, Jonathan D., Claudio Lucarelli, and Christopher A. Powers. 2015. "Paying Attention or Paying Too Much in Medicare Part D." *American Economic Review* 105(1): 204–33.

Kling, Jeffrey R., Sendhil Mullainathan, Eldar Shafir, Lee C. Vermeulen, and Marian V. Wrobel.

2012. "Comparison Friction: Experimental Evidence from Medicare Drug Plans." *Quarterly Journal of Economics* 127(1): 199–235.

Kőszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association* 4(4): 673–707.

Kőszegi, Botond, and Adam Szeidl. 2012. "A Model of Focusing in Economic Choice." *Quarterly Journal of Economics* 128(1): 53–104.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112(2): 443–78.

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin. 2003. "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics* 118(4): 1209–48.

Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives* 25(3): 17–38.

Madrian, Brigitte C., and Dennis F. Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *Quarterly Journal of Economics* 116(4): 1149–87.

Malmendier, Ulrike, and Young Han Lee. 2011. "The Bidder's Curse." *American Economic Review* 101(2): 749–87.

Matějka, Filip, and Alisdair McKay. 2015. "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." *American Economic Review* 105(1): 272–98.

McCall, John Joseph. 1970. "Economics of Information and Job Search." *Quarterly Journal of Economics* 81(1): 113–26.

Mullainathan, Sendhil, Joshua Schwartzstein, and William J. Congdon. 2012. "A Reduced-Form Approach to Behavioral Public Finance." *Annual Review of Economics* 4: 511–40.

O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review* 89(1): 103–24.

Ortoleva, Pietro. 2013. "The Price of Flexibility: Towards a Theory of Thinking Aversion." *Journal of Economic Theory* 148(3): 903–34.

Oster, Emily, Ira Shoulson, and E. Ray Dorsey. 2013. "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease." *American Economic Review* 103(2): 804–30.

Pauly, Mark V., and Fredric E. Blavin. 2008. "Moral Hazard in Insurance, Value-Based Cost Sharing, and the Benefits of Blissful Ignorance." *Journal of Health Economics* 27(6): 1407–17.

Rees-Jones, Alex, and Dmitry Taubinsky. 2016. "Heuristic Perspectives of the Income Tax: Evidence and Implications for Debiasing." NBER Working Paper 22884.

Rosenfield, Donald B., and Roy D. Shapiro. 1981. "Optimal Adaptive Price Search." *Journal of Economic Theory* 25(1): 1–20.

Rothschild, Michael. 1974. "Searching for the Lowest Price When the Distribution of Prices is Unknown." *Journal of Political Economy* 82(4): 689–711.

Schneider, Henry S. 2016. "The Bidder's Curse: Comment." *American Economic Review* 106(4): 1182–94.

Schwartzstein, Joshua. 2014. "Selective Attention and Learning." *Journal of the European Economic Association* 12(6): 1423–52.

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50(3): 665–90.

Stigler, George J. 1961. "The Economics of Information." *Journal of Political Economy* 69(3): 213–25.

Taubinsky, Dmitry, and Alex Rees-Jones. Forthcoming. "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment." *Review of Economic Studies.*

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* Penguin Books.

Woodford, Michael. 2012. "Inattentive Valuation and Reference-Dependent Choice." May 2. http://www.columbia.edu/~mw2230/InattentiveValue%20Harvard%20Seminar.pdf.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 179–194

Do Economists Swing for the Fences after Tenure?

Jonathan Brogaard, Joseph Engelberg, and Edward Van Wesep

Tenure is pervasive in American higher education: every one of the top 500 colleges and universities in the United States as ranked by *US News and World Report* has some kind of tenure-granting system. The "philosophical birth cry" of the academic tenure system (Metzger 1973) was the 1915 statement of the American Association of University Professors (AAUP). Formalized in the 1940 Statement of Principles on Academic Freedom and Tenure (available at https://www.aaup.org/report/1940-statement-principles-academic-freedom-and-tenure), a joint statement of the AAUP and the Association of American Colleges (AAC) proclaimed: "Tenure is a means to certain ends; specifically: (1) freedom of teaching and research and of extramural activities, and (2) a sufficient degree of economic security to make the profession attractive to men and women of ability. Freedom and economic security, hence, tenure, are indispensable to the success of an institution in fulfilling its obligations to its students and society."

It is clear why associations of professors favor the intellectual freedom and economic security provided by the institution of tenure. The benefits of tenure could also be more philosophical: academic freedom in teaching and research is important for reasons other than the generation of highly cited papers. But for economists, it is natural to ask a more specific question: Under what conditions is

■ *Jonathan Brogaard is Assistant Professor of Finance, Foster School of Business, University of Washington, Seattle, Washington. Joseph Engelberg is Professor of Finance and Accounting, Rady School of Management, University of California–San Diego, La Jolla, California. Edward Van Wesep is Associate Professor of Finance, Leeds School of Business, University of Colorado, Boulder, Colorado. Their email addresses are brogaard@uw.edu, jengelberg@ucsd.edu, and Edward.VanWesep@Colorado.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.1.179

doi=10.1257/jep.32.1.179

tenure part of an optimal contract? After all, the incentives provided by the threat of termination are perhaps the starkest incentives faced by most employees, and tenure removes those incentives.

A variety of additional justifications for tenure have been proposed (for a discussion in this journal, see McPherson and Shapiro 1999). For example, the carrot of tenure can incentivize effort pre-tenure, allow for lower salaries, induce more selective hiring, or attract risk-averse but talented individuals to academia. As one example of prominent work in this area, Ito and Kahn (1986) argue that tenure-style assurances of the possibility of long-lived employment—not only in academia, but also in civil service, law and accounting firms, and a number of other workplaces—can be viewed as an efficient method of risk-sharing when an employer wants an employee to make a risky human capital investment. Other reasons for tenure can arise due to peculiarities in the nature of academia. For example, professors are the members of a university best able to identify talented prospective hires, and without tenure, they might fear losing their jobs if they hire too well (Carmichael 1988; see also Friebel and Raith 2004; Siow 1998). In addition, tenure, which both protects senior faculty from dismissal and makes them residual claimants on any rents in the institution, gives senior faculty the incentive to monitor university leadership (Brown 1997).

Finally, society may benefit more from research that is truly groundbreaking than research which is more incremental. Trying to do something innovative and failing looks a lot like shirking, so motivating risky innovation may require the assurance of tenure (Manso 2011).[1] Our focus is on this last argument: do academics respond to receiving tenure by attempting more ground-breaking "home run" research and in this way "swinging for the fences"?

In order to answer this question, we hand-collect a sample of all academics who pass through economics or finance departments at top 50 US schools from 1996 through 2014. From this sample of over 2,000 faculty, we consider two variables in the years before and after each academic receives tenure: the total number of publications and the number of "home run" publications. The number of publications is a measure of the quantity of output; the number of home run publications focuses on *highly influential* output and is a measure of the quality of output.

We find that both variables have values that peak at tenure and decline thereafter. The average number of annual publications falls by approximately 30 percent over the two years after tenure is granted and falls by an additional 15 percent over the subsequent eight years. The average number of annual home run publications also falls by 30 percent over the two years following tenure, but falls by an additional 35 percent over the subsequent eight years. Combining these facts, we find that not

[1] Some additional theories relevant to academic tenure include the discussion of "up-or-out" employment settings, where workers either receive a promotion or are let go at some stage. Kahn and Huberman (1988) examine employers with "up-or-out" promotion practices in a situation of two-sided uncertainty and moral hazard, while Waldman (1990) emphasizes the role of signaling in up-or-out settings. Demougin and Siow (1994) consider careers within hierarchies, and the conditions under which firms will prefer to promote from within.

only do both the overall publication rate and the home run rate fall, but the likelihood of a given publication being a home run falls by approximately 25 percent during the ten years following tenure. Conversely, papers in the bottom 10 percent of citations are actually published more frequently in the years following tenure than in the tenure year.

These patterns suggest two insights. First, the fall in publication rates over the two years following tenure is consistent with the notion that tenure tends to be granted when publication success has been achieved, and so a degree of reversion to the mean is expected. The timing of tenure is at least in part endogenous: faculty can advance early if they are highly productive early in their careers, and they can switch employers if they are unlikely to get tenure at a first institution. Further, the timing of publication is endogenous: faculty can time their efforts on various projects to maximize the number of publications before their tenure clock expires.

A second insight, more relevant for our paper, is that publication behavior from years two through ten after tenure suggests that after receiving tenure, economics faculty reduce risk-taking and the quality of their output falls. This might occur in a number of ways: adding coauthors; advertising new papers less at conferences and seminars; working on easier topics, which can be published in good journals but have less impact; or any number of other behaviors. We consider several alternatives to our explanation—increased nonresearch service work post-tenure or an increased tendency of tenured researchers to branch out into new subject areas—and show that none can fully explain what we find.

This paper does not evaluate the broad and multidimensional case for and against tenure. But it does suggest that at least for economists, tenure is not providing incentives to undertake research in the same quantity and quality that led up to the tenure decision.

Quantity and Quality of Research: Pre- and Post-Tenure

To construct our sample, we hand-collected employment and publishing data among economics and finance professors. We began by including all faculty who were employed at any of the top 50 economics or finance departments in the United States in any year from 1996 through 2014. This process involved use of the Wayback Machine (waybackmachine.org) and hand-collection of curriculum vitae (CVs). We collected a total of 2,763 names, 2,092 of whom are eventually granted tenure at some point prior to 2014. After collecting the set of faculty and their tenure years, we match this database to a database of publications and citations for 51 leading economics and finance journals. More detail, including the list of journals, is provided in the online appendix available with this paper at http://e-jep.org.

Quantity of Research

We begin by evaluating a subset of faculty who are present in our data for at least five years prior to their tenure year and ten years after. We require pre- and

post-tenure data for all faculty. We exclude two other groups, which are people who would mechanically strengthen the increase in observed publication rates prior to tenure and the decrease post-tenure. First, some faculty were granted tenure less than five years from their first appearance in our sample. This was usually because they began their careers at government agencies, where they may not have been expected to publish to the same degree as in academia. Second, some faculty left academia prior to, or less than ten years after, tenure, or received tenure after 2004, and thus were unlikely to publish as often post-tenure. Including these faculty would severely bias downward pre-tenure publication rates relative to post-tenure rates (especially if we were to include faculty who never receive tenure). We therefore drop them.

The final dataset contains 980 faculty, all of whom received tenure prior to 2004. To address the issue of coauthored papers, we define an author's contribution to a publication as $1/N$, where N is the number of authors on the publication. We show in the online appendix that if we do not adjust for coauthorship our qualitative results remain unchanged.

Figure 1 presents the per-capita author-adjusted number of papers published by this subset relative to the year that the academic was first tenured. The year marked "tenure" is the first year in which the researcher was tenured, the year marked "–1" is the year before, and so on. The figure shows annual publications increasing monotonically prior to tenure, peaking in the neighborhood of the granting of tenure and declining steadily thereafter.

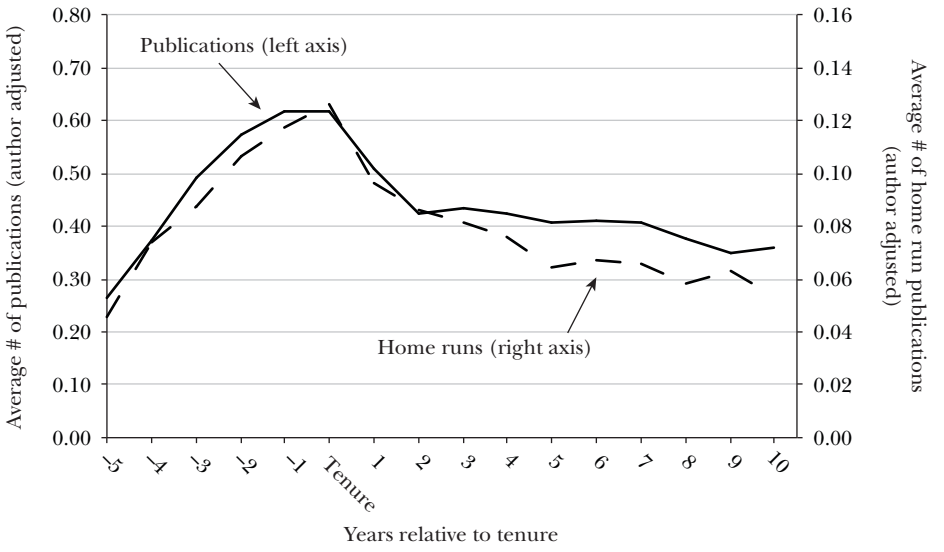
In order to interpret magnitudes, note that the height of the solid line in year –2 is 0.57. This means that our 980 researchers produced, on average, 0.57 author-adjusted papers in solid journals in the year prior to the one in which they were put up for tenure. This number would imply $0.57 \times 980 = 559$ solo-authored papers, $0.57 \times 2 \times 980 = 1{,}117$ dual-authored papers, or higher numbers of three-or-more authored papers. In fact, the number averages across these types of papers.

Home Runs

We also calculate the number of home run publications, defined as publications that, as of 2015, were among the 10 percent most cited of all papers published in a given year. The plot of the number of home runs shown in Figure 1 is largely similar to the plot of publications, peaking in the tenure year and falling thereafter. The number of home runs is anywhere from 1/7 to 1/5 of the number of publications. These numbers are greater than 10 percent for two reasons. First, the faculty in our sample are mostly associated with prestigious departments and presumably publish more-cited papers. Second, we only include economists who get tenure, and these are likely more cited as well.

We can calculate the ratio of the two plots in Figure 1: home run publications divided by all publications. The series is noisy, but clearly exhibits a substantial decrease over the period two to ten years post-tenure. This decrease will be important in teasing out potential explanations for the patterns we see.

Figure 1
Publications and Home Runs around Tenure



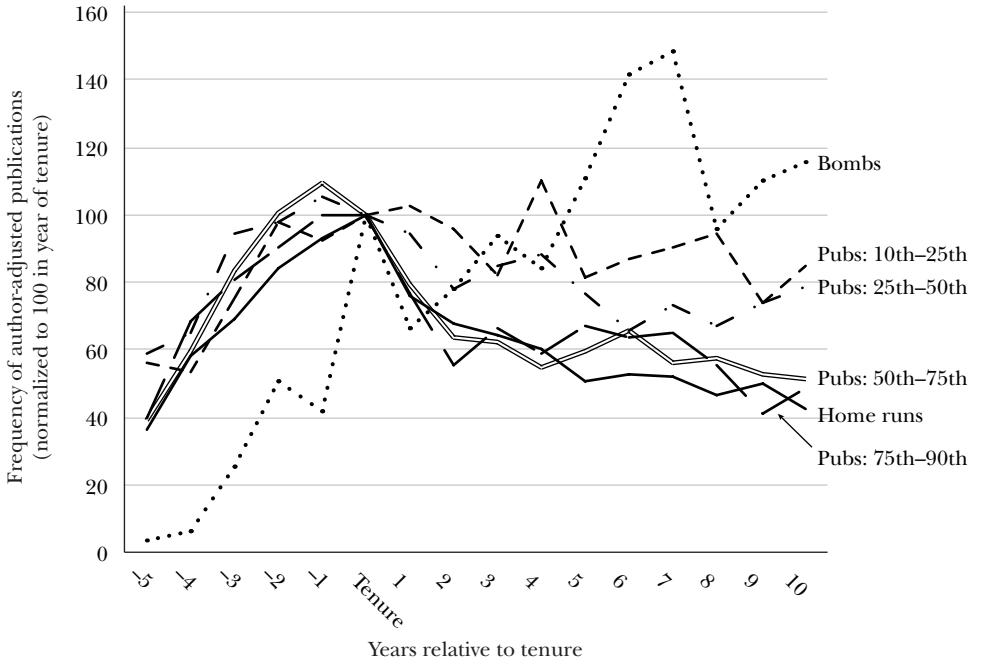
Note: This figure plots the number of publications and the number of those publications that were "home runs" in event time, where the event is tenure. A publication in an economics or finance journal is defined as a home run if it has more citations than 90 percent of all economics and finance publications appearing in the same year. The sample consists of 980 faculty whose publication activity we observe for at least 5 years before tenure and 10 years following tenure. Each author on a publication is credited with the inverse of the number of authors on the publication (for example, an article with four authors counts as .25 of a publication for each author). "Average # of publications (author adjusted)" is the sum of the cohort's publications divided by 980. "Average # of home run publications (author adjusted)" is the sum of the cohort's home runs divided by 980. The height of each curve therefore represents the average number of publications (left axis) and home run publications (right axis) for a member of our sample in each year of his or her career, measured from the year of tenure, where this author only receives $1/N$ credit on an N-authored paper.

Together, these facts provide suggestive evidence that tenure is associated with peak academic production, in terms of the quantity of publications, the quantity of home run publications, and the likelihood that a given publication becomes a home run.

A Closer Look at Risk-Taking in Economic Research

In order to gain additional perspective on risk-taking in publication, we also plot the rate of non–home-run publications. In Figure 2, we assign each paper a category based on its citations: if it was in the top 10 percent of citations for papers published in that year, it is called a "home run"; if it is in the lowest 10 percent, it is called a "bomb." We split papers further into 10th to 25th, 25th to 50th, 50th to 75th, and 75th to 90th percentile groups. We calculate the number of papers published by authors from five years before to ten years after tenure. We calculate the number of papers in each citation bucket in the year of tenure, and normalize that value to 100.

Figure 2
Publications around Tenure by Citation Percentile



Note: This figure plots the relative frequency of author-adjusted publications by citation percentile around tenure. Home runs (Bombs) are those above the 90th (below the 10th) percentile of all economics and finance papers in the same year. "Pubs: 75th–90th," "Pubs: 50th–75th," "Pubs: 25th–50th," and "Pubs: 10th–25th" are similarly defined. Each series is normalized to 100 in the year of tenure.

Figure 2 shows that publication rates of all paper types increase in tandem up to the year of tenure, but there is substantial divergence afterward. The one category of paper with consistently higher quantity post-tenure is the "bombs." Indeed, ten years after tenure, the most common category (the highest value) is bombs, the second-highest is publications in the 10–25th percentile, and so on down to the least-common category of home runs.

Table 1 shows the results of regressions, which provide a sense of the statistical significance of the changes in overall and home run publication rates shown in Figure 1. We estimate variants of the following linear model:

$$Pub_{i,t} = \alpha + \beta_t + \gamma_i + \delta_{i,t}^{-5,-1} I_{i,t}^{-5,-1} + \delta_{i,t}^{+1,+5} I_{i,t}^{+1,+5} + \delta_{i,t}^{+6,+10} I_{i,t}^{+6,+10} + \varepsilon_{i,t},$$

where β_t is a year fixed effect designed to capture differential publication rates over time; γ_i is a researcher fixed effect designed to capture differential publication rates across researchers; $I_{i,t}^{m,n}$ is a dummy variable taking a value of 1 if, in year t, researcher i is between m and n years from tenure (with positive values of m and n

Table 1
Publications and Home Runs around Tenure

	Dependent variable					
	Publications	*Publications*	*Publications*	*Home runs*	*Home runs*	*Home runs*
Years -5 to -1 (pre-tenure)	–0.155*** (0.028)	–0.168*** (0.028)	–0.092*** (0.027)	–0.040*** (0.112)	–0.046*** (0.011)	–0.032*** (0.011)
Years +1 to +5 (post-tenure)	–0.178*** (0.028)	–0.153*** (0.028)	–0.226*** (0.027)	–0.046*** (0.011)	–0.038*** (0.011)	–0.052*** (0.011)
Years +6 to +10 (post-tenure)	–0.237*** (0.028)	–0.186*** (0.028)	–0.373*** (0.037)	–0.065*** (0.011)	–0.049*** (0.011)	–0.085*** (0.015)
Observations	15,680	15,680	15,680	15,680	15,680	15,680
Year fixed effects	NO	YES	YES	NO	YES	YES
Researcher fixed effects	NO	NO	YES	NO	NO	YES
R^2	0.0072	0.0233	0.2780	0.0037	0.0161	0.2164
p-value for test: Years +1 to +5 = years +6 to +10	0.0000	0.0167	0.0000	0.0000	0.0260	0.0000

Note: The dependent variable in the first (last) three columns is author-adjusted publications (home runs). Years -5 to -1, Years +1 to +5, and Years +6 to +10 are the 5 years before tenure, the first 5 years after tenure, and the next 5 years after tenure, respectively. The final row reports the *p*-value from a linear restriction test, which tests the equality of coefficients on Years +1 to +5 and Years +6 to +10. Ordinary least squares standard errors are in parentheses.
*, **, and *** indicate statistical significance at the 10, 5, and 1 percent levels, respectively.

representing post-tenure dates and negative values representing pre-tenure dates) and zero otherwise; and $\delta_{i,t}^{m,n}$ is the coefficient on the tenure time dummy variable associated with years m to n after tenure.

The excluded year for any researcher i is the tenure year, so all coefficients are average publication rates relative to a professor's tenure year. Depending on the regression, $Pub_{i,t}$ may represent the overall author-adjusted number of publications, or the author-adjusted number of home run publications for researcher i in year t.

In Table 1, column 1, we perform the analysis with no year or researcher fixed effects. That is, this regression ignores the facts that publication rates have increased over time and that some authors publish more than others. On average, 0.155 fewer author-adjusted publications occur in the five years prior to tenure, 0.178 fewer in the five years after, and 0.237 fewer in the five years following that. Publications are lower before and after tenure, and even lower the longer after tenure one goes.

In Table 1, column 2, we add year fixed effects to account for the fact that publication rates have increased over time, and in column 3 we add year and researcher fixed effects. The inclusion of year fixed effects increases the R^2 but has a relatively small effect on the coefficients: relative to the first column, the publications before tenure are a little lower and those after tenure are a little higher, but the peak at tenure remains. The inclusion of researcher fixed effects, however, has a substantial effect. It's no surprise that the R^2 is again higher; by design, researcher fixed effects

will absorb variation in across-researcher publication rates. However, the coefficients shift in a way that strengthens the peak at tenure and the decline post-tenure.

In Table 1, columns 4–6, we repeat the analyses of the first three columns using data on home run papers and find substantial reductions in the rate at which authors produce home run papers, in periods both before and after tenure. As in columns 1–3, the number of home runs produced decreases in the five years following tenure and continues to decrease in the five years after that.

We test whether we can statistically differentiate the coefficients on the dummy variables for the periods one to five years and six to ten years post-tenure. We perform a Wald F-test for the equality of the coefficients on the dummy variables for years +1 to +5 and years +6 to +10. In all six cases, we strongly reject the null hypothesis that the coefficients are equal. Not only do the rates of publications and home runs fall in the five years following tenure, but they continue to fall in the five years after that.

Summarizing Patterns

In sum, we have shown that: 1) Publication and home run rates rise to tenure, peaking in the year a researcher comes up for tenure and a researcher's first year as tenured faculty. 2) Publication and home run rates fall markedly in the two years following tenure. 3) Publication and home run rates fall by 15 and 35 percent, respectively, from years two through ten after tenure, while bomb rates increase by 35 percent.

Our interpretation of these facts is: 1) Junior faculty get better at publishing in their first few years, and publication lags are long, leading to an increase in the publication rate of all paper qualities as tenure approaches. 2) Tenure is typically granted when success is achieved. Because of publication lags, this leads to high publication rates in the year that the researcher is coming up for tenure as well as during the following year. 3) As tenured faculty age, there is a decade-long decline in the production of publications and home runs and an increase in the production of bombs. We believe that the most consistent explanation for these two declines is a change in risk-taking by academic researchers.

Alternative Explanations for Productivity Declines Post-Tenure

In this section, we consider five alternative explanations for the patterns shown thus far that could help to explain our findings. We will show that none can fully explain the patterns that we see. While they all may be at work, a reduction in risk-taking by academic researchers seems to be relevant as well.

Perhaps This Is a "Time since PhD" Effect

A number of studies have shown that research productivity follows a hump-shape over age, first rising and then falling (for example, Oster and Hammermesh 1998; Levin and Stephan 1989; Gingras, Larivière, Macaluso, and Robitaille 2008).

This could be because aging directly affects the ability of an academic to produce top-rate research, or it may be because the marginal effect of an additional top publication on an academic's professional outcome decreases as the number of publications increases. These studies have not, however, looked specifically at the timing of tenure. It could be a coincidence that the various factors that lead to the rise and fall of academic productivity over a lifetime just happen to peak at the year tenure was granted. Can we separate out a specific effect of tenure in our data?

To investigate this possibility, we split the sample by the year in which a researcher was granted tenure: fifth year, sixth year, and so on. Naturally, the sample in each case is substantially smaller than for the full sample, adding noise to our plots, so we make several adjustments to boost the sample (details and plots can be found in the appendix).

For those tenured in five years, the year of peak production of both papers and home runs is the tenure year. For those tenured in six years, the publication rate is highest in the year before tenure and the tenure year; the home run publication rate peaks in the tenure year and the year after. For those tenured in seven years, both publications and home runs peak in the year the candidate is up for tenure. As the data become noisier (fewer people are tenured each year after seven), the peaks are less clear but the general shape persists: people publish more and better papers in the run-up to tenure and fewer after.

These patterns suggest that it is not simply aging that is causing the patterns observed in Figures 1 and 2. The year of tenure itself is special, not just the number of years since graduate school.

Perhaps It's the Rise in Service, Teaching, and Nonacademic Obligations Post-Tenure

It is possible—even likely—that many faculty in our sample experience increased expectations of university service after tenure, including advising, department chairing, serving as a dean, and other administrative and committee member responsibilities. Indeed, these additional administrative and service responsibilities are one of the aforementioned justifications for tenure, and thus generally consistent with a tenure-based explanation for the data. Also, tenured faculty often have more opportunities for outside opportunities after tenure, like consulting or book-writing. These factors tend to reduce publication rates, even if the researcher's aggregate effort over all activities increases post-tenure.

To investigate this explanation for our findings, we return to Figure 2. Suppose that authors have some ability to distinguish between projects likely to be successful and those likely to fail. A researcher who experiences an increase in nonresearch obligations post-tenure would presumably seek to reduce effort on low-impact projects. Thus, one might expect the number of publications to fall, but the number of home runs to remain similar and the share of home runs to rise. We do not see this result.

Instead, the likelihood that a given publication becomes a home run falls from 20 percent the tenure year, to 15 percent ten years later. This decline is substantial. Moreover, this reduction is not due solely to mean reversion and the endogenous

timing of tenure. The decline begins in earnest three years after the tenure year, which is four years after the researcher is up for tenure. Any papers that led the researcher to get tenure would likely have been published before then.

We can also point to Figure 2 to show that the production of bombs actually rises in the ten years following the granting of tenure. Service obligations should not drive an *increase* in the production of low-citation papers! These patterns suggest that while nonresearch post-tenure obligations may affect productivity, there is more to the story.

Perhaps Tenure Encourages Researchers to Branch Out

Tenure may not lead to an increase in home run publications, but it may lead to an increase in interdisciplinary work, which may take time and perhaps not lead to papers with high citations counts but still help ideas to germinate in important ways. There are several ways in which branching out could appear in our data: choosing new coauthors, publishing in new journals, and publishing in new areas.

To consider this possibility, consider the set of faculty in our dataset who eventually received tenure and for whom we can observe their first 15 years in academia. Then, for those 15 years, we estimate variants of the following linear probability model, in which each observation is a single publication:

$$Y_{i,t,r,s} = \alpha + \beta_t + \gamma_r + \delta_s + T_r + X_i + \varepsilon_{i,t,r,s}$$

where $Y_{i,t,r,s}$ is a dummy variable measuring whether paper i written at time t by researcher r, who has been a professor for s years, represents branching out (defined in three ways below); β_t is a year fixed effect designed to capture differential tendencies to branch out over time; γ_r is a researcher fixed effect designed to capture differential tendencies to branch out across researchers; δ_s is an event-time fixed effect designed to measure different tendencies to branch out as a researcher ages; and T_r is a dummy variable indicating whether the researcher has tenure.

In the first regression in Table 2, a paper is defined as branching out (that is, $Y_{i,t,r,s} = 1$) if it involves a new coauthor. In this case, X_i represents a "coauthor count" fixed effect, which accounts for the fact that researchers with more prior coauthors tend to add new coauthors more rarely.[2]

In our second regression, a paper is defined as branching out if it is published in a journal in which the researcher has never before published. For example, if a researcher has published only in finance journals and then publishes a new paper in the *Journal of Labor Economics*, then $Y_{i,t,r,s} = 1$ for this paper. In this case, X_i represents a "prior journal count" fixed effect, which accounts for the fact that it is more difficult to publish in a new journal when one has already published in many different journals previously.

[2] A coauthor count fixed effect is actually a set of fixed effects. The first takes a value of one if the author has never had a coauthor on any of her prior papers, and zero otherwise. The second takes a value of one if the author has only ever worked with one other coauthor previously, and zero otherwise, and so forth.

Table 2
Other Forms of Risk-taking

	Dependent variable:					
	New Coauthor (1)	*New Journal (2)*	*New Area (3)*	*New Coauthor (4)*	*New Journal (5)*	*New Area (6)*
Tenure	−0.006	−0.053***	−0.041***	0.005	0.012	0.009
	(0.011)	(0.011)	(0.008)	(0.014)	(0.014)	(0.010)
Observations	16,280	16,280	16,280	16,280	16,280	16,280
Year fixed effects	YES	YES	YES	YES	YES	YES
Event year fixed effects	YES	YES	YES	YES	YES	YES
Coauthor count fixed effects	YES	NO	NO	YES	NO	NO
Journal count fixed effects	NO	YES	NO	NO	YES	NO
Area count fixed effects	NO	NO	YES	NO	NO	YES
Researcher fixed effects	NO	NO	NO	YES	YES	YES
R^2	0.060	0.131	0.098	0.235	0.272	0.307

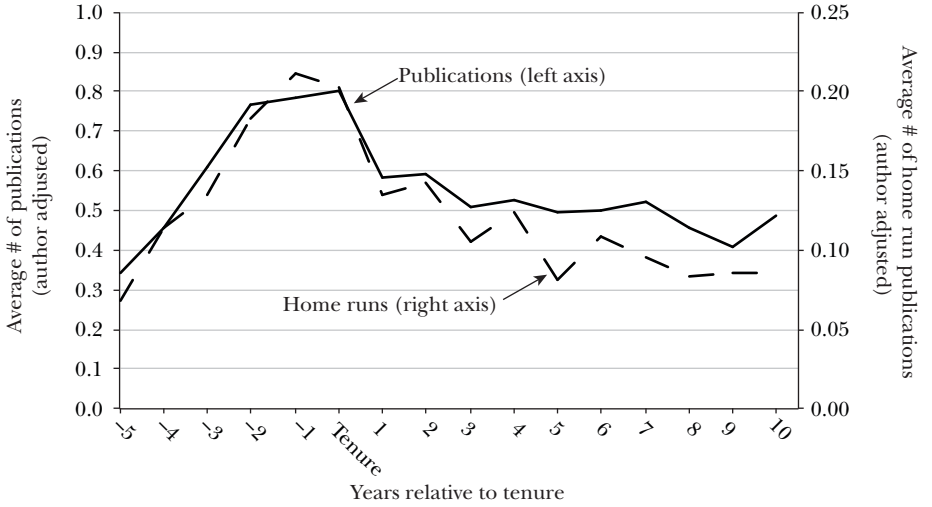
Note: Each observation is a professors' publication. The dependent variable in the first and fourth columns is a dummy variable equal to one if the publication is with a new coauthor and zero otherwise. In the second and fifth columns, it is a dummy variable equal to one if the publication is in a new journal for the professor, and in the third and sixth columns, it is a dummy variable equal to one if it is in a new subject area for the professor. Subject areas are grouped into accounting, econometrics, finance, general interest, industrial organization, international economics, labor economics, law and economics, macroeconomics, microeconomics, monetary economics, and public economics. Each professor's first publication is excluded (because "new" is trivially equal to one). Ordinary least squares standard errors are in parentheses.
*, **, and *** indicate statistical significance at the 10 percent, 5 percent, and 1 percent, respectively.

In our third regression, a paper is defined as branching out if it is published in a new subject matter area. For example, the *Journal of Labor Economics* is defined as being in the area of labor economics, whereas the *Journal of Financial Economics* is defined as being in the area of finance. General interest journals are more difficult to categorize, so we define them to be in their own area.[3] In this case, X_i represents a "prior areas count" fixed effect, which accounts for the fact that it is more difficult to publish in a new area when one has already published in many areas previously.

Results are displayed in Table 2. In the first three columns, we do not include researcher fixed effects, and in the last three we do. Based on regressions in columns 1 and 4 there do not appear to be substantial differences in the tendency to add new coauthors pre- and post-tenure. Based on regressions in columns 2 and 5, it does not appear to be more common for researchers to publish in new journals post-tenure. If anything, regression 2 suggests a weak tendency to publish in new journals less often. Importantly, this is not because tenured faculty have been out longer; this is accounted for with event-time fixed effects. It is also not because tenured faculty have already published in more journals, making it harder to publish in a new one; this

[3] Interested readers can find the assignment of journals to areas in Table 2a in the online Appendix.

Figure 3
Publications and Home Runs around Tenure in Elite Schools



Note: This figure plots the number of publications and the number of those publications that were "home runs" in event time, where the event is tenure. A publication in an economics or finance journal is defined as a home run if it has more citations than 90 percent of all economics and finance publications appearing in the same year. The sample consists of 333 faculty whose publication activity we observe for at least five years before tenure and ten years following tenure, and who initially placed at one of the following schools: University of California–Berkeley, University of Chicago, Columbia University, Harvard University, Massachusetts Institute of Technology, Northwestern University, University of Pennsylvania, Princeton University, Stanford University, and Yale University. Each author on a publication is credited with the inverse of the number of authors on the publication (for example, an article with four authors counts as .25 of a publication for each author). "Average # of publications (author adjusted)" is the sum of the cohort's publications divided by 333. "Average # of home-run publications (author adjusted)" is the sum of the cohort's home runs divided by 333. The height of each curve therefore represents the average number of publications (left axis) and home run publications (right axis) for a member of our sample in each year of his or her career where the author only receives $1/N$ credit on an N-authored paper.

is accounted for with journal count fixed effects. Finally, based on regressions 3 and 6, there is no evidence that tenured faculty branch out more by publishing in a new area. If anything, there is weak evidence of a tendency to branch out less.

Risk-Taking May Decline on Average, But Perhaps Not for Elite Faculty

The preceding results are averages. Perhaps faculty at the most prestigious departments, who produce the lion's share of truly influential papers, exhibit a different pattern of publication after tenure.

In Figure 3, we perform the same analysis as in Figure 1, and plot publications and home runs for five years pre- to ten years post-tenure, but restrict the sample to faculty who begin their careers at a subset of particularly prestigious schools: University of California–Berkeley, University of Chicago, Columbia University, Harvard University, Massachusetts Institute of Technology, Northwestern University, University of Pennsylvania, Princeton University, Stanford University, and Yale University.

As in each of our subsamples thus far, publications and home runs peak in the year the researcher is up for tenure and in the first year of tenure. Both fall markedly in the first two years post-tenure and then consistently from years two through ten post-tenure. The smaller sample size means that there is more noise than for the full sample, but the pattern is striking. Faculty who begin their careers at elite schools have the same publication pattern as those who begin elsewhere. Indeed, from years two through ten post-tenure, the drop in the publication rate is 15 percent and the drop in the home run rate is 35 percent—precisely the same as in the full sample. The patterns we identify are present for faculty at both higher- and lower-ranked schools.

Perhaps It Takes Time for Truly Novel Research to Gain Traction

Perhaps truly influential papers take time to become known and cited. Perhaps Manso (2011) is correct in suggesting that the type of innovation for which tenure seeks to provide incentives is precisely the riskier type, which may take more time to catch on. To analyze whether this is the case, we restrict our sample to faculty who were tenured by 1994, and therefore papers published no later than 2004. As we evaluate the citations as of 2014, this allows at least ten years for a paper to catch on. As in other subsample analyses, there is more noise, but the pattern is still present. In fact, we once again see a 15 percent reduction in the publication rate, and a 35 percent reduction in the home run rate, in years two through ten post-tenure. The persistence of these ratios is surprisingly stable.

Perhaps This is True Only for Faculty at Schools with Poor Post-Tenure Contracting

It may be the case that some schools employ contracting techniques that encourage their faculty to swing for the fences, but the positive outcomes at these schools are outweighed by faculty at schools with poor contracting. We cannot observe the quality of contracting at every school at our sample, but one natural decomposition would be to separate public and private universities in the United States. Public institutions are subject to a variety of laws governing the compensation, hiring, and retention of state employees. Private institutions are largely free to design compensation programs at will.

If we split the sample into those faculty first tenured at US private schools and those first tenured at US public schools (and drop those first tenured elsewhere), we find that, at public schools, publication rates fall 36 percent in the two years following tenure, and a further 16 percent, in the subsequent eight years. At private schools, publication rates fall 30 percent in the two years following tenure, and a further 12 percent in the subsequent eight years. If we focus on home runs, the publication rate at public schools falls 43 percent in the two years following tenure, and a further 44 percent in the subsequent eight years. At private schools, the home run rate falls 25 percent in the two years following tenure and a further 32 percent in the subsequent eight years. If we focus on the likelihood that a publication becomes a home run, at public schools it falls by 11 percent in the two years following tenure and a further 33 percent in the subsequent eight years. At private

schools, the likelihood that a publication becomes a home run actually rises by 7 percent in the two years following tenure but then falls 22 percent in the subsequent eight years.

In sum, the patterns are similar at public and private schools. Publication rates fall by similar amounts at both types of school, but home run rates fall substantially less at private schools, providing weak evidence that contracting might help schools avoid this problem. (Figures and additional discussion on these points can be found in the online Appendix.)

Perhaps Our Definition of Home Run Is Too Generous

In our sample, approximately 1/7 of papers become home runs. Are there that many papers that are truly impactful? Perhaps researchers are publishing fewer above-average papers, but really are producing more spectacular papers.

Choosing a cutoff for home runs is a balancing act. Increasing the cutoff selects for papers that are more influential, but it reduces the number of papers defined as home runs and thus injects noise. We choose the top 10 percent as our threshold as a balance between the objectives of accurately measuring influence and minimizing noise. But perhaps we choose incorrectly.

We therefore consider an alternative to the home run, which we call the grand slam. A paper is defined to be a grand slam if it is in the top 5 percent of all papers published in its publication year, measured by citations as of 2014. We find that the rate of grand slams is approximately half of the rate of home runs, which is to be expected, and the pattern is similar. The rate of grand slam publications falls by 29 percent in the two years following tenure, and falls a further 32 percent in the subsequent eight years. These numbers closely align to those for home runs. (Again, further discussion and the associated figure can be found in the online Appendix.)

Conclusion

This paper should not be read as an indictment of the institution of tenure. As noted at the start of the paper, tenure has an array of costs and benefits. In this paper, we consider only one aspect of tenure, and only for researchers in economics and finance. However, focusing on that one aspect, it does not appear that academic economists respond to the greater professional and intellectual freedom that tenure should provide by sustaining their earlier research effort or by taking the chances that lead to more home run research. Among academic economists at research-oriented institutions, rates of publication and home run publications rise up to the year of tenure and fall for a decade thereafter.

From one point of view, our paper contributes to a small empirical literature on the effect of tenure on academic output. Holley (1977) evaluates the productivity, both in terms of quantity and quality, of 97 sociologists surrounding their tenure dates. He finds decreased performance on both dimensions post-tenure. Li and Ou-Yang (2010) focus on economics and finance faculty from the top 25 schools

and find no statistically significant difference in impact pre- and post-tenure. The difference between their result and ours seems due to the substantial increase in statistical power that we achieve by including more faculty from more schools and a wider set of journals. Yoon (2016) analyzes the publication and citation rates for US law school professors and finds that those rates rise to tenure and fall slightly thereafter. He analyzes only the first ten years of a professor's career, little of which is post-tenure, so he cannot separate the effect of endogenous timing of tenure from the longer-run effects on productivity or effort.

We also believe that our findings raise some practical questions for academic economists and their institutions. For economists, the findings suggest that they should be wary of allocating their research time in a way that seems likely to lead to low-impact papers, and instead consider if there is a way for them to continue their earlier research efforts—at least in terms of quality, if not necessarily in quantity. When making a tenure decision, departments of economics and their home institutions should be aware that the research productivity of the person receiving tenure is likely to decline, in both quantity and quality terms, over the following decade. Thus, institutions should consider whether there are methods to sustain (or at least not to impede) high-quality research efforts.

■ *We thank seminar participants at Australia National University, University of Colorado Boulder, and Colorado State University. We also thank conference participants at a number of meetings: the California Corporate Finance Conference; the Labor and Finance Group Winter Meeting; the Jackson Hole Finance Conference; the Financial Institutions, Regulation and Corporate Governance Conference; the Pacific Northwest Finance Conference; and the Utah Winter Finance Conference. Helpful additional comments were provided by Renee Adams, Jonathan Berk, Peter DeMarzo, Vincent Gregoire, Lubos Pastor, and Luigi Zingales. We also are grateful for research support from Kothai Priyadharshini Alagarsamy, Jonathan Bannick, Hadley Evarts, Rui Han, Tarun Patel, Ryan Skorupski, Maia Szafer, and Che Zhang.*

References

Brown, William O., Jr. 1997. "University Governance and Academic Tenure: A Property Rights Explanation." *Journal of Institutional and Theoretical Economics* 153(3): 441–61.

Carmichael, Lorne H. 1988. "Incentives in Academics: Why Is There Tenure?" *Journal of Political Economy* 96(3): 453–72.

Demougin, Dominique, and Aloysius Siow. 1994. "Careers in Ongoing Hierarchies." *American Economic Review* 84(5): 1261–77.

Friebel, Guido, and Michael Raith. 2004. "Abuse of Authority and Hierarchical Communication." *RAND Journal of Economics* 35(2): 224–44.

Gingras, Yves, Vincent Larivière, Benoît Macaluso, and Jean-Pierre Robitaille. 2008. "The Effects of Aging on Researchers' Publication and Citation Patterns." *PLoS ONE* 3(12): e4048.

Holley, John W. 1977. "Tenure and Research

Productivity." *Research in Higher Education* 6(2): 181–92.

Ito, Takatoshi, and Charles Kahn. 1986. "Why Is There Tenure?" Center for Economic Research Discussion Paper 228.

Kahn, Charles, and Gur Huberman. 1988. "Two-Sided Uncertainty and 'Up-or-Out' Contracts." *Journal of Labor Economics* 6(4): 423–44.

Levin, Sharon G., and Paula E. Stephan. 1989. "Age and Research Productivity of Academic Scientists." *Research in Higher Education* 30(5): 531–49.

Li, Si, and Hui Ou-Yang. 2010. "Explicit Incentives, Implicit Incentives, and Performance: Evidence from Academic Tenure." Available at SSRN: https://ssrn.com/abstract=399240.

Manso, Gustavo. 2011. "Motivating Innovation." *Journal of Finance* 66(5): 1823–60.

McPherson, Michael S., and Morton O. Shapiro. 1999. "Tenure Issues in Higher Education." *Journal of Economic Perspectives* 13(1): 85–98.

Metzger, Walter P. 1973. "Academic Tenure in America: A Historical Essay." *Faculty Tenure: A Report and Recommendations* by the Commission on Academic Tenure in Higher Education, pp. 93, 135–36. San Francisco: Jossey-Bass.

Oster, Sharon M., and Daniel S. Hamermesh. 1998. "Aging and Productivity among Economists." *Review of Economics and Statistics* 80(1): 154–56.

Siow, Aloysius. 1998. "Tenure and Other Unusual Personnel Practices in Academia." *Journal of Law, Economics, & Organization* 14(1): 152–73.

Waldman, Michael. 1990. "Up-or-Out Contracts: A Signaling Perspective." *Journal of Labor Economics* 8(2): 230–50.

Yoon, Albert. 2016. "Academic Tenure." *Journal of Empirical Legal Studies* 13(3): 428–53.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 195–210

Retrospectives
Cost-Push and Demand-Pull Inflation: Milton Friedman and the "Cruel Dilemma"

Johannes A. Schwarzer

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please contact Joseph Persky, Professor of Economics, University of Illinois, Chicago, at jpersky@uic.edu.

Introduction

James Tobin (1967) spoke for a substantial share of the economics profession at the time when he described the Phillips curve as a "cruel dilemma," because it suggested that full employment was not compatible with price stability. Many economists of the 1950s and 1960s regarded inflation not as an exclusively monetary demand-pull phenomenon, but as also emerging due to cost-push forces related to market institutions and imperfections, like strong unions, which interacted with monetary policy and aggregate demand. In his famous presidential address to the American Economic Association in 1967, Milton Friedman (1968) presented an analytical framework to support his long-held position that no such structural conflict between the two policy goals existed and that monetary policy was not only an inappropriate but also ineffective tool to influence the rate of unemployment in the long run. Friedman's criticism regarding the Phillips curve trade-off built

■ *Johannes A. Schwarzer is a Lecturer in Economics, University of Hohenheim, Stuttgart, Germany. His email address is schwarzer.econ@gmail.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.32.1.195

doi=10.1257/jep.32.1.195

on two pillars: First, his framework defined a natural rate of unemployment that would result from the structures and institutions of a real-world economy, including factors cited as cost-push forces such as union power. Second, Friedman emphasized the role of inflation expectations in the Phillips curve, so that a trade-off between unemployment and inflation could only exist in the short run before inflation expectations fully adjusted—but in the long run, the economy would revert to its natural rate of unemployment.

The following discussion begins by focusing on the importance of cost-push factors that many economists emphasized with respect to Phillips curve analysis in the 1950s and 1960s. I then turn to the evolution of Friedman's thought on this issue. His arguments through the 1950s and into the 1960s grappled explicitly with the notion that inflation might have an underlying cost-push dimension, though Friedman rejected the idea of structural cost-push inflation particularly due to union power. In Friedman's (1968) presidential address, factors cited as cost-push forces like unions become determinants of the natural rate of unemployment and as such are rendered irrelevant for the inflationary process by his analytical framework, while fully-adjusting inflation expectations become a decisive element for monetary policy to consider. Friedman's critics argued that he was dodging the issue by equating his concept of the natural rate of unemployment with full employment when these ideas need not be the same. Moreover, critics made a case that ongoing cost-push inflation could exist at full employment and therefore a genuine Phillips curve dilemma cannot be swept aside by assumption.

Though Friedman's rejection of cost-push inflation is one of the pillars of his criticism of the Phillips curve trade-off, his presidential address is mainly remembered today (together with the parallel work of Phelps 1967, 1968) for pointing out the role of inflation expectations in macroeconomic analysis, and for distinguishing that an economy would adjust to its natural rate of unemployment in the long run but could display an unemployment-inflation trade-off in the short run. These ideas have played a large role in the macroeconomic research that followed. However, questions about what causes inflation to move—and in recent years, what has caused inflation to remain so stable—have continued to the present day.

The "Cruel Dilemma" and the Phillips Curve

The "cruel dilemma" view of the Phillips curve was based on earlier experience that inflation sometimes emerged before full employment was achieved. As one example, Morton (1950, p. 26) points at 1937, when wages rose despite millions of unemployed in the United States. Some episodes after World War II, particularly the years from 1955 to 1958, also featured a rise of inflation to what seemed like high levels at the time, despite ongoing unemployment. Thus, the policy issue at hand was "that inflation may exist concurrently with non-frictional unemployment" (Bowen 1960, p. 205).

A number of economists discussed the possibilities that general market imperfections such as bottlenecks and factor immobility could lead to inflation even without full employment, but the usual focus was on trade unions. In 1950, US union membership had risen to 40 percent of the labor force outside of agriculture (Slichter 1954, p. 329). Though the issue of cost-push inflation due to unions was already the focus of prominent economists before World War II (Humphrey 1977), there was no consensus on how to explain the wage-setting behavior of unions. Some argued that unions acted as monopolies (for example, Friedman 1951b, p. 206), while others were skeptical of applying that framework to union behavior (for example, Haberler 1951, pp. 34–35, n. 2). More fundamentally, the question arose as to whether union behavior could be understood as maximizing the income of its members or if union behavior is driven by political aspects (Reder 1952). Despite these disputes, there was a general consensus that union wage demands also pulled up nonunion wages, which caused the impression that "our wage-fixing arrangements have an inflationary bias" (Slichter 1954, p. 345). In the same vein, Slichter (1952, p. 54) pointed at the inflationary effects of strong unions even before full employment is achieved: "At some point short of full employment the bargaining power of most unions becomes so great that they are able to push up money wages faster than the engineers and managers can increase output per man-hour." In the context of the UK economy, *The Economist* wrote a series about "The Uneasy Triangle" (1952a, b, c) and remarked that there seems to be a "three-cornered incompatibility between a stable price level, full employment, and the free collective bargaining."

In the contemporary editions of Paul Samuelson's prominent introductory textbook (1958), he emphasized that this kind of cost-push inflation is at the heart of the issue of macroeconomic policy debates:

> It is hardly too much to say that this price-wage question is the biggest unsolved economic problem of our time: *Can business, labor, and agriculture learn to act in such a way as to avoid inflation whenever private or public spending brings us anywhere near to full employment?* A wage and price policy for full employment—that is America's greatest problem and challenge (p. 360).

At the end of the 1950s, the original Phillips (1958) curve paper seemed to provide a quantitative answer to the inflation–unemployment problem because the results (p. 299) implied that it would be possible to stabilize the price level in the United Kingdom with an unemployment rate of 2.5 percent.[1] However, when Samuelson and Solow (1960, p. 192) estimated a Phillips curve for American data, their results suggested that price stability would require an unemployment rate of 5 to 6 percent, which was regarded as too high a cost to accept for price stability

[1] Forder (2014, forthcoming) questions the views that the Phillips curve was originally seen as promoting inflation, and that Friedman (1968) was intended as a challenge to the feasibility of such policy. In Schwarzer (2016, pp. 113ff.), I critically consider these and related issues.

by a substantial share of economists (according to a survey of "economic experts at colleges and universities" by the Joint Economic Committee 1958). Indeed, at this time a 3 percent unemployment rate was often associated with "full employment" (for example, Bronfenbrenner and Holzman 1963, p. 627), which implied 4 to 5 percent of inflation based on the Samuelson–Solow Phillips curve and as such conflicted with the goal of price stability.

Although the Phillips curve was originally interpreted as a demand-pull relation (Schwarzer 2012, p. 982), it was in principle compatible with cost-push approaches (Lipsey 1960, p. 31) and thus became a handy framework within which to discuss inflation from either source. As a prominent example, the 1961 edition of Samuelson's textbook (p. 383) interpreted the downward-sloping Phillips curve as "a modified cost-push model" and added: "There is, so to speak, a choice for society between reasonably high employment with maximal growth and a price creep, or reasonably stable prices with considerable unemployment; and it is a difficult social dilemma to decide what compromises to make."

The concern about the risk of inflation without apparent general excess demand, often phrased as a result of dynamics arising from union wage-bargaining, persisted through the 1960s and beyond. For example, it was discussed in contemporary studies aimed at policy advice such as the reports of the Commission on Money and Credit (1961, pp. 15ff.) or the Council of Economic Advisers (1966, pp. 178ff.). Gardner Ackley (1966, pp. 70–71), who chaired the Council of Economic Advisers under the Johnson administration, wrote that the "tendency of wage rates to increase every year, no matter what" is to be regarded as an "institutional inflationary bias." In a similar vein, Solow (1966, p. 42) pointed out that the tendency of money wages and prices to rise while there is still slack in the economy "creates a dilemma for public policy."

None of the possible solutions to this inflation–unemployment dilemma had much appeal for a variety of economic, political, or social reasons.

For example, one policy option was to accept an ongoing positive rate of inflation. However, this was thought to result in undesirable side-effects such as the distortion of saving–investment decisions or the slowing down of growth (for a discussion, see Schwarzer 2014, pp. 187–88). Moreover, it was often feared, as Jacoby (1957, p. 23) warned, that "[w]hat began as 'creeping' inflation will become 'running' inflation." Therefore, Jacoby concluded, "[t]he policy of a responsible government *must* be to maintain an absolutely stable price level; it is a dangerous illusion to think otherwise."[2] Indeed, there was a strong aversion towards inflation in general as, for example, Clark (1960, p. 12) remarked that an inflation rate of 2.5 percent "would be quite serious enough, and materially higher rates would spell economic calamity."

A contrasting option was to "do business with the [inflation] dragon—buying some reduction in the degree of inflation by feeding him a certain number of jobs" (Lerner 1967, p. 3). However, this solution, that is, "[t]he creation of

[2]Such concerns over the stability of a positive rate of inflation were not unanimous at the time. For a more optimistic discussion also framed by the apparent policy dilemma, see Lipsey (1961).

unemployment as a cure for inflation," as many economists feared, "is politically unacceptable" (Smithies 1957, p. 281). Of course, the Phillips curve also offered in-between choices, with Reuber (1962) providing one of the first detailed analyses, albeit focused on the Canadian economy.

Other options seemed no more attractive. Solow (1966, p. 43) pointed out that any remedy that involves breaking the market power of unions or large firms was "more than a little unrealistic." On the other hand, "direct price and wage controls," as remarked by Samuelson (1958, p. 359), "would involve a degree of planning incompatible with past, and probably present, philosophical beliefs of the great majority of the American people." A common proposal, often viewed as a compromise, was to restrain inflation through a voluntary incomes policy of following wage and price "guideposts" (in the phrasing of the Council of Economic Advisers 1962, pp. 185ff.). These guideposts suggested that wages rise in line with trend productivity growth while prices should follow unit labor costs, "so that expansion policy could close the [output] gap and not be dissipated in price increases" (Staff of the Cabinet Committee on Price Stability 1969, p. 125). There was ongoing controversy over whether such a program would have beneficial effects—or whether a voluntary program would have any effect at all. For some, the Phillips curve encapsulated this issue of cost-push inflation and the possible role of guideposts. A few months before Friedman's presidential address, Samuelson (in Burns and Samuelson 1967, pp. 54–55) emphasized its relevance, stating that the Phillips curve "is one of the most important concepts of our times" so that "[a]ny criticism of the guideposts which does not explicitly take into account the Phillips curve concept I have to treat as having missed the fundamental point of all economic policy discussions." Indeed, as I will show in the next section, Friedman's criticism of cost-push inflation became embedded into the Phillips curve framework in his presidential address.

How Friedman's Views Evolved

Milton Friedman had long argued that there was no structural conflict between price stability and full employment, or as stated in his presidential address, no long-run trade-off between unemployment and inflation for monetary policy.[3] As an early example of his views, Boulding (1951, p. 79) summarized in rhyme the results of a discussion taking place during a 1950 conference about the economic effects of unions: "We all (or nearly all) consent/ If wages rise by ten per cent/ It puts a choice before the nation/ Of unemployment or inflation." The one economist at that 1950 conference not joining the consensus view, and thus the "nearly all" referred to in Boulding's verse, was Milton Friedman. A few years later, when asked by the Joint Economic Committee about his view on the conflict between inflation

[3]This section benefited from Ed Nelson's comments on a previous draft of the paper and his comprehensive contributions on Friedman's work. See Nelson (forthcoming, pp. 586ff.) for an in-depth analysis of how Friedman's views on cost-push inflation evolved over time.

and unemployment, Friedman (in Joint Economic Committee 1959, p. 626) clearly stated that no dilemma existed:

> Senator Bush. ... One of the principal objectives of this committee's work this year is to try to find out the relationship between the maintenance of employment and price stability. ... Do you think those are mutually conflicting or not? ...
>
> Mr. Friedman. I do not believe they are mutually conflicting. …

The underlying assumption behind this view that there is no structural conflict between full employment and price stability can be found in Friedman's (1963 [1968], p. 39; 1966b, p. 18) famous statement: "Inflation is always and everywhere a monetary phenomenon." In this view, ongoing price increases cannot be due to cost-push pressures but are the outcome of demand-pull forces driven by monetary policy.

But in the years leading up to Friedman's 1967 presidential address, he did on various occasions acknowledge the possibility of cost-push inflation arising from collective bargaining as well as certain contexts in which an unemployment–inflation trade-off might arise. For example, in the 1950 conference on the role of unions, Friedman (1951a, pp. 243–44) mentioned "the logical possibility of inflation from the cost side in an economy of strongly organized producer groups," so that "the phenomenon of higher prices plus unemployment ... is logically possible" but—at least in the USA—not "an empirically important possibility" (see also Friedman 1951b, pp. 227–28; 1955, p. 404).

Friedman also suggested at times that inflation could arise if the monetary authority feels responsible for achieving full employment, if this desire for full employment implies accommodating any wage increase, no matter how large. For example, Friedman (1963[1968], pp. 29–30) writes that "it is true that the upward push in wages produced inflation, not because it was necessarily inflationary but because it happened to be the mechanism which forced an increase in the stock of money," which is why "[f]ull employment policy is ... a modern invention for producing inflation." In Friedman's (p. 39) view, this happened in "Britain these past few years."

This line of argument suggests the possibility of a policy dilemma in which high union wage demands force a policymaker to decide between unemployment and inflation. In Friedman's view, this still means that monetary policy is ultimately responsible for whether inflation occurs. But as the next section will discuss in more detail, Friedman's contemporary critics often saw his argument as an evasion of structural cost-push pressures that should also be treated along with demand-pull factors as a cause of inflation. Indeed, those concerned about cost-push inflation often argued that monetary policy is likely to be driven by such cost-push pressures (Bronfenbrenner 1950, pp. 622–23) or that the effective money supply (via the expansion of bank credit or an increase in velocity) would rise endogenously in the

wake of cost-push pressures (Machlup 1960, p. 127; Fleming 1961, p. 515). In pure cost-push scenarios—that is, if cost-push pressures are completely independent of the rate of unemployment and actual output—no such accommodation of the cost-push implied a one-for-one fall of income to compensate the rise in the price level, while the Phillips curve at least offered the option for the policymaker to moderate cost-push pressures by reducing aggregate demand.

In the year before Friedman's presidential address in December 1967, he confronted the argument of cost-push pressures from unionization in a more direct way. In a discussion reprinted in a 1966 conference volume, Friedman (1966a, p. 57) reasoned that any level of market power of unions is in line with price stability since "[i]nsofar as market power has anything to do with possible inflation, what is important is not the *level* of market power, but whether market power is *growing* or not." The reasoning is that a one-time cost-push inflation[4] due to a growing market power of unions is possible, as unions exploit that increase in market power to establish "the maximum real income and real wage rate that they thought it was worth their while." But once that increase in market power is fully exploited and the higher wage established, there will be no further push for even higher wages.

In the same comment, Friedman (1966a, p. 60) combines this rejection of cost-push theories of inflation with the importance of fully-adjusting inflation expectations in an explanation of why guideposts (in addition to concerns such as the likely distortion of the price system as discussed in Friedman 1966b, pp. 37–38) are not an appropriate answer to inflation:

> Hence, the alleged case for the guidelines seems to me to rest on two basic fallacies: first, that market power is a source of rising prices, and second—on the belief that somehow or other you can fool the people all the time—that by increasing the rate of monetary expansion, you can thereby induce people to maintain a [permanently] lower level of unemployment.

Also in the same comment, Friedman (1966a, p. 60) offered a definition of the natural rate of unemployment: "But for any given labor market structure, there is some natural level of unemployment at which *real* wages would have a tendency to behave in accordance with productivity." Notice that Friedman's definition takes the structure of the labor market as given, and in this way suggests that the natural rate of unemployment might well be different between two countries with high and low rates of unionization. Furthermore, real wages at the natural rate of unemployment grow in line with productivity by definition, so that the absence of cost-push wage-pressure is an inherent feature of the natural rate concept.

[4] This one-time rise of the price level due to the increase in union power is, in Friedman's view, not necessarily to be interpreted as cost-push but as demand-pull inflation even without any increase in either the money supply or its velocity, since the increase in union power will reduce potential output (Nelson forthcoming, pp. 76, 414, 591).

As becomes clear from the very first paragraph of Friedman's (1968) AEA presidential address, his talk is shaped by this ongoing debate of whether or not the goals of "high employment, stable prices, and rapid growth" are "mutually compatible." As Friedman (in Taylor 2001, p. 124) later recalled, a basic cornerstone of his presidential address, the natural rate hypothesis, "grew out of the discussions about [income] guidelines and, in particular, out of the Samuelson and Solow paper on the Phillips curve." His address tied together many of these themes and made the arguments explicit in a highly visible setting, but also refocused the arguments in ways that would prove of lasting salience in macroeconomic research. In Friedman's criticism (1966a) of the Phillips curve trade-off in the year before his presidential address, the explicit rejection of cost-push inflation goes hand in hand with the important role of fully-adjusting inflation expectations. In the address, his criticism regarding cost-push inflation is now fully translated and embedded into the natural rate concept, making his rejection of cost-push inflation an integral part of his framework, though less visible than the emphasis on the role of fully-adjusting inflation expectations.

The natural rate of unemployment in Friedman's (1968, p. 8) address is determined by "the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on." In this way, labor unions and other factors cited as cost-push forces are incorporated into the natural rate: for example, as Friedman (p. 9) writes, "the strength of labor unions ... make[s] the natural rate of unemployment higher than it would otherwise be."[5] Given Friedman's (1966a) earlier reasoning that unions at constant market power can at best only be made responsible for high but not continuously rising wages, treating the strength of labor unions as a determinant of the natural rate, and therefore rendering them irrelevant for the inflationary process, follows naturally. Because all other cost-push factors that were discussed as having the potential to build up inflationary pressure also become determinants of the natural rate, only monetary forces are left for explaining inflation, so that the natural rate "separate[s] the real forces from monetary forces" (Friedman 1968, p. 9). Indeed, unemployment below this natural rate is labeled "excess demand for labor" (p. 8), which hints at the demand-pull view and suggests the coincidence of full employment with the natural rate. Because the natural rate of unemployment is compatible with price stability as well as with any rate of inflation or even deflation, there is no necessity to choose between the two policy objectives.[6] Furthermore, even if such a conflict existed, there would be no possibility for monetary policy to

[5] See also Friedman (1972, p. 194; 1975, p. 30). In his Nobel Lecture (Friedman 1977, p. 458), the strength of labor unions is not explicitly listed as a determinant of the natural rate, though "the extent of competition or monopoly" is (see also Friedman 1966a, p. 60).

[6] In the same year when he gave his presidential address, Friedman (1967, p. 13) explicitly stated that "[w]e do not have to choose between inflation and unemployment." A few years later, Friedman (1975, p. 14) made his view clear that at the natural rate, "[u]nemployment is zero—which is to say, as measured, equal to 'frictional' or 'transitional' unemployment."

"peg the rate of unemployment for more than very limited periods" (p. 5) anywhere else than at the natural rate. With inflation expectations ultimately coinciding with actual inflation and having a unit weight in the Phillips curve, the Phillips curve becomes vertical in the long run with only "unanticipated inflation" (p. 11) altering the rate of unemployment in the short run.

In short, because there is no need *and* no possibility to choose between the two policy objectives, monetary policy should and can only focus on the desired nominal target such as the rate of inflation without any connection to real objectives such as unemployment in the long run (p. 11).

Reactions to Friedman: Cost-Push and Demand-Pull Entangled

Many economists at the time interpreted Friedman's (1968) reasoning regarding the "cruel dilemma" not as innovative, but as dodging the issue. The counterargument was that Friedman, by subsuming all kinds of market imperfections and cost-push forces under his definition of the natural rate of unemployment, was defining away the conflict between full employment and price stability. In response to Friedman's (1966a, b) essays in the run-up to the presidential address, Ackley (1966, p. 68) expressed his "complete disagreement with Mr. Friedman's proposition that in any operationally meaningful sense inflation is caused by an excessive increase in the quantity of money and by nothing else." Though Ackley does not deny that inflation can be the result of general excess demand, he emphasizes that "the definition of productive capacity, by comparison with which total demand may be excessive, is itself a significant issue" and makes an implicit reference to Friedman's natural rate concept:

> I believe the evidence is inescapable that we can have inflation without what *I* would call excess demand, as the result of excessive income claims by labor or business or both. Of course, one can define this possibility out of existence. If one defines the total productive capacity of the economy as that degree of utilization which, if exceeded, leads to rising prices, then all inflation becomes excess demand inflation and the issue disappears.

From this perspective, the issue was that Friedman's natural rate concept offered no solution to the perceived policy dilemma, since accepting structural cost-push elements such as union power as a determinant of and limit to the full employment level implied giving up on the original full employment target, and instead regarding any further inflation as caused purely by demand-pull factors for which restrictive monetary policy was an appropriate and, in effect, costless solution.

Other critics focused on Friedman's argument that cost-push inflation is only reasonable if there is a change in market power. Haberler (1969, p. 69–70) emphasized the difference between monopolies and labor unions, since the latter "are out for large *annual* wage increases and not merely for a once-for-all substitution

of a higher monopoly wage for the lower competitive wage." The reasoning was that even without unions, real wages would rise with productivity, and money wages would rise at price stability and therefore render "it a perfectly natural objective for union policy to push continuously for money wage increases that are higher than is compatible with full employment equilibrium at stable prices." Thus, Haberler (p. 71) remarked that "once labor unions have acquired strength . . . we can expect continuing wage push without any further acquisition of 'market power.'" Haberler (1972, p. 238) hence emphasized that "[w]age-push by powerful labor unions is an obvious reality" and complained that "[n]o more would need to be said about the existence of the problems, if some monetarists had not denied the connection between inflation and the monopoly power of labor unions for so long." With respect to the theory of monopolies, Ackley (1966, p. 71) emphasized that it is market power as such, and not necessarily a rise in market power, that is important. An increase in demand that strengthens a producer's ability to realize the desired monopoly price would in its wake increase costs for other producers, who would also raise prices in order to restore their desired margins. Given a general nominal downward inflexibility of prices and wages due to market power on both sides (as argued by a report from Ackley's Council of Economic Advisers 1966, p. 179), inflation would arise, which would further be fueled by desired wage adjustments on the side of labor to make up for the rise in the cost of living. As such, an inflationary spiral may be possible without any additional rise in market power.

Four years after Friedman's address, James Tobin (1972), in his own presidential address to the American Economic Association on the subject of "Inflation and Unemployment," revisited what he had earlier called the "cruel dilemma." In contrast to previous critics of Friedman, Tobin (p. 14) endorsed the argument that market power of unions cannot be a source of ongoing cost-push inflation and thus implicitly accepted one pillar of Friedman's argument. Nonetheless, Tobin cautioned that the natural rate should not be unconditionally equated with full employment (p. 2), and he still argued in favor of a genuine long-run Phillips curve trade-off. Tobin reasoned that when there are downward nominal rigidities, ongoing relative price adjustments necessary to remove sectoral disequilibria can be a source of inflationary pressure without general excess demand (pp. 9ff.). This "passive cost-inflation mechanism" (as it was called in Dow 1962, p. 45) was regarded by many economists as another important source of the perceived incompatibility of full employment and stable prices, and thus served as a rationale for accepting a positive rate of inflation as the outcome of a full employment economy subject to permanent change and growth creating ongoing sectoral disequilibria (Schwarzer 2016, pp. 125ff.).

Friedman acknowledges the prevalence of nominal rigidities throughout his writings (as discussed in Nelson 2008, pp. 103ff.) but in his presidential address instead turns that into an argument for the merits of a stable overall price level. Friedman (1968, p. 13) argues that "in the United States, there is only a limited amount of flexibility in prices and wages. We need to conserve this flexibility to achieve changes in relative prices and wages that are required to adjust to dynamic

changes in tastes and technology. We should not dissipate it simply to achieve changes in the absolute level of prices that serve no economic function." Friedman thus rejected the alleged long-run benefit of inflation for facilitating relative price adjustments because it may eliminate the (downward) flexibility of prices and wages. In this context, he emphasized that the best monetary policy can do is to assure "that the average level of prices will behave in a known way in the future—preferably that it will be highly stable."

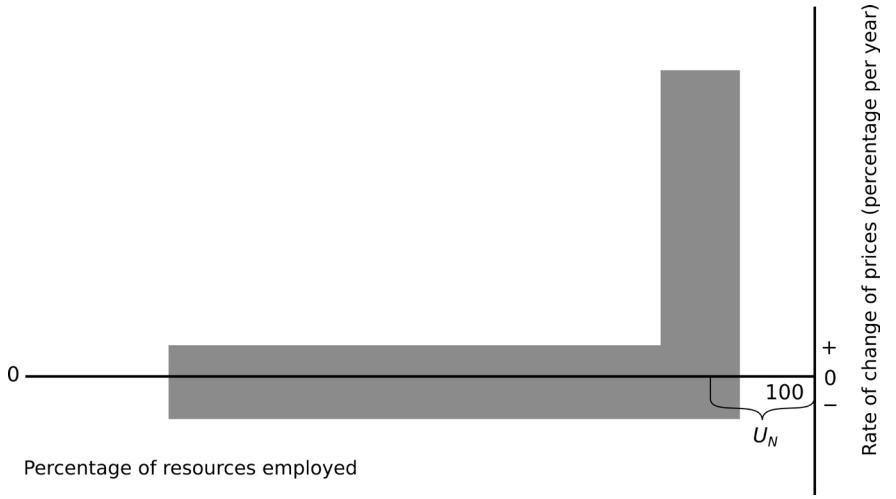
These professional disputes also lingered regarding whether there are theoretical arguments that inflation expectations do not always fully adjust or are not fully translated into wages and prices (for example, Tobin 1972, p. 13) and whether such full adjustment can be found in the data (for example, Solow 1969). In the aftermath of Friedman's (1968) speech, prominent textbooks started to comment on the role of inflation expectations and the natural rate, but nonetheless continued for some years to teach both cost-push and demand-pull factors and to discuss a conventional downward-sloping Phillips curve which offered choices for policymakers.

For example, the 1970 edition of Samuelson's introductory textbook (p. 811, n. 10, figure 41-3) includes side-by-side diagrams of a Phillips curve for the pure forms of cost-push (horizontal line at the rate of cost-push inflation) and demand-pull (vertical line at full employment) inflation and discusses their policy implications, with pure cost-push as "no tradeoff possible" and with pure demand-pull as "no tradeoff being necessary." However, the downward-sloping Phillips curve which combines both horizontal and vertical forces is presented as "a dramatic way of describing the dilemma for macro policy" (p. 811) because "[i]f we move leftward toward full employment, before we get there, wages and prices may tend to rise and keep rising" (p. 810, caption of figure 41-2).

In an alternative textbook approach, Lipsey (1975, p. 804) did not choose to illustrate the difference between short-run and long-run Phillips curves which is implicitly outlined in Friedman's presidential address and explicitly argued in Phelps (1967, 1968). Instead, Lipsey focused on the implications of Friedman's assumptions about the inflationary process, as shown in Figure 1. On the one hand, because the Phillips curve is vertical at the natural rate of unemployment, while there is otherwise no tendency for inflation to become positive until the natural rate is reached, pure demand-pull inflation is assumed in Lipsey's interpretation of Friedman, so that Lipsey (pp. 803–804) speaks of "[a] revival of the L-shaped relation" and of "orthodox demand-pull theory." However, in contrast to the original L-shaped curve in which full employment is at an utilization rate of 100 percent (pp. 800–801), this rate at which prices start to rise is now shifted to the left and thus lower, so that Lipsey (p. 804, caption of figure 51.7) speaks of "[t]he new theory of the L-shaped relation with a non-zero natural rate of unemployment (U_N)."

Lipsey (1975) presented Friedman's (1968) approach within the concept of the conflict-free demand-pull-only L-shaped relation, while a corresponding downward-sloping Phillips curve, in line with Tobin's (1972) reasoning of ongoing market disequilibria, still implies a conflict between the two policy objectives of full employment and price stability (Lipsey 1975, p. 803). Thus, Lipsey's interpretation

Figure 1
The Natural Rate as an L-Shaped Supply Curve Concept.



Source: Reproduced (redrawn and modified in order to deliver a better print quality) from *An Introduction to Positive Economics*, p. 804 by Richard G. Lipsey, Fourth Edition, 1975, published by Weidenfeld and Nicolson, with permission from The Orion Publishing Group, London. © 1963 by Richard G. Lipsey.

of Friedman's natural rate framework encapsulates the sentiment that Friedman was dodging the issue by offering a different inflationary process and by equating the natural rate with full employment.

Cost-push forces as a source of inflation were still discussed and prominent in the United States after Friedman's (1968) presidential address. Along with the other examples given here, Arthur Burns, who was appointed Federal Reserve Chairman in January 1970, began to endorse a cost-push view of inflation, while Friedman continued his criticism of cost-push inflation and his opposition to guideposts (as discussed in Nelson 2007, pp. 154ff.; Nelson and Schwartz 2008, pp. 841ff.).

Conclusion

From the 1950s into the 1970s, many economists argued that cost-push forces and in particular the market power of unions played an important role in explaining how inflation could arise even when an economy had not reached full employment, as illustrated by the Phillips curve trade-off between price stability and full employment. As Tobin (1967, p. 102) noted in the short paper that emphasized the "cruel dilemma," inflation "is neither demand-pull nor cost-push, or, rather, it is both" so that "[t]he Phillips curve approach forces us to confront squarely the fact that our goal[s] for prices and employment are not wholly reconcilable." Friedman, on the other hand, argued that structural cost-push inflation in the sense of an inflationary

bias at full employment is not realistic, since only growing market power makes union-induced cost-push inflation theoretically feasible. In Friedman's (1968) presidential address, factors cited as cost-push forces such as union power hence become determinants of the natural rate of unemployment, while the structural rate of inflation solely depends on the path of monetary policy. In sum, Friedman's "view is optimistic, because it means that there is no long-run conflict between high employment and price stability" (Friedman 1972, p. 194). This "modern doctrine" (Nelson 2009, p. F345) regarding the inflationary process, as well as Friedman's emphasis on the full adjustment of inflation expectations, have played a major role in macroeconomic research ever since and continue to shape monetary policy.

However, questions about the determinants of inflation have resurfaced in recent years. These questions have focused on the "inflation puzzle" of why inflation has been so stable, despite seemingly large shifts like the Great Recession and the dramatic expansionary monetary policies in its wake (for a comprehensive assessment, see Miles, Panizza, Reis, and Ubide 2017). In a series of speeches, Federal Reserve Chair Janet Yellen (2016, 2017a, b) highlighted three important elements to be analyzed further for an understanding of inflation in recent times: the concept and estimation of the natural rate of unemployment (also stressed by Phelps 2017); the role and measurement of inflation expectations; and the specification of the underlying framework for analyzing inflation dynamics. The answers to such questions will be sought in the ways that demand-pull and cost-push factors interact in an economy with adjusting inflation expectations, imperfect markets, and nominal rigidities. In these arguments, the distinctions and controversies surrounding Friedman's presidential address of 50 years ago may well play a central role.

■ *Preliminary versions of this paper were presented at the PhD Seminar of the German Keynes Society in Darmstadt, Germany, February 17–18, 2014, and at the 17th Summer School on History of Economic Thought, Economic Philosophy, and Economic History with the topic "Unemployment and the Social Question" in Zaragoza, Spain, September 1–7, 2014. I am grateful to Joseph Persky for his encouragement to submit the paper. I thank James Forder, Niels Geiger, Harald Hagemann, Thomas Humphrey, David Laidler, Richard Lipsey, Arash Molavi Vasséi, Edward Nelson, Jean-Pierre Potier, and André Straus for most valuable comments and suggestions. Earlier correspondence with Ronald Bodkin, Grant Reuber, and Robert Solow helped to spark some core ideas presented in this paper. I am indebted to Gordon Hanson and Mark Gertler for helpful remarks and to Timothy Taylor for his invaluable assistance in shaping the paper into its final form.*

References

Ackley, Gardner. 1966. "The Contribution of Guidelines." In *Guidelines, Informal Controls, and the Market Place: Policy Choices in a Full Employment Economy,* edited by George P. Shultz and Robert Z. Aliber, 67–78. University of Chicago Press.

Boulding, Kenneth. E. 1951. "Selections from the Discussion of the Clark and Haberler Papers." In *The Impact of the Union,* edited by David McCord Wright, 63–79. New York: Harcourt, Brace and Company.

Bowen, William G. 1960. "'Cost Inflation' versus 'Demand Inflation': A Useful Distinction?" *Southern Economic Journal* 26(3): 199–206.

Bronfenbrenner, Martin. 1950. "Trade Unionism, Full Employment, and Inflation: Comment." *American Economic Review* 40(4): 622–24.

Bronfenbrenner, Martin, and Franklyn D. Holzman. 1963. "Survey of Inflation Theory." *American Economic Review* 53(4): 593–661.

Burns, Arthur F., and Paul A. Samuelson. 1967. *Full Employment, Guideposts and Economic Stability*. Washington, DC: American Enterprise Institute for Public Policy Research.

Clark, John M. 1960. *The Wage-Price Problem*. New York: Committee for Economic Growth without Inflation, The American Bankers Association.

Commission on Money and Credit. 1961. *Money and Credit: Their Influence on Jobs, Prices, and Growth—The Report of the Commission on Money and Credit.* Englewood Cliffs, NJ: Prentice-Hall.

Council of Economic Advisers. 1962. *The Annual Report of the Council of Economic Advisers.* In *Economic Report of the President together with The Annual Report of the Council of Economic Advisors,* transmitted to the Congress January 1962. Washington, DC: United States Government Printing Office.

Council of Economic Advisers. 1966. *The Annual Report of the Council of Economic Advisers.* In *Economic Report of the President together with The Annual Report of the Council of Economic Advisors,* transmitted to the Congress January 1966. Washington, DC: United States Government Printing Office.

Dow, J. C. R. 1962. "Internal Factors Causing and Propagating Inflation: II." In *Inflation—Proceedings of a Conference held by the International Economic Association,* 37–53, edited by Douglas C. Hague. London and New York: Macmillan & St. Martin's Press.

Economist, The. 1952a. "The Uneasy Triangle." August 9 (5685): 322–23.

Economist, The. 1952b. "The Uneasy Triangle—II: The Limits of Free Bargaining." August 16 (5686): 376–78.

Economist, The. 1952c. "The Uneasy Triangle—III: How Full Employment?" August 23(5687): 434–35.

Fleming, Miles. 1961. "Cost-Induced Inflation and the Quantity Theory of Money." *Economic Journal* 71(283): 512–20.

Forder, James. 2014. *Macroeconomics and the Phillips Curve Myth*. Oxford Studies in the History of Economics. Oxford University Press.

Forder, James. Forthcoming. "What was the Message of Friedman's *Presidential Address* to the American Economic Association?" *Cambridge Journal of Economics.*

Friedman, Milton. 1951a. "Selections from the Discussion of Friedman's Paper." In *The Impact of the Union,* edited by David McCord Wright, 235–59. New York: Harcourt, Brace and Company.

Friedman, Milton. 1951b. "Some Comments on the Significance of Labor Unions for Economic Policy." In *The Impact of the Union,* edited by David McCord Wright, 204–34. New York: Harcourt, Brace and Company.

Friedman, Milton. 1955. "Marshall and Friedman on Union Strength: Comment." *Review of Economics and Statistics* 37(4): 401–406.

Friedman, Milton. 1963 [1968]. "Inflation: Causes and Consequences." In *Dollars and Deficits: Inflation, Monetary Policy and the Balance of Payments*, edited by Milton Friedman, 21–71. Englewood Cliffs, NJ: Prentice-Hall. (First published 1963. Bombay: Asia Publishing House for the Council for Economic Education.)

Friedman, Milton. 1966a. "Comments." In *Guidelines, Informal Controls, and the Market Place: Policy Choices in a Full Employment Economy,* edited by George P. Shultz and Robert Z. Aliber, 55–61. University of Chicago Press.

Friedman, Milton. 1966b. "What Price Guideposts?" In *Guidelines, Informal Controls, and the Market Place: Policy Choices in a Full Employment Economy,* edited by George P. Shultz and Robert Z. Aliber, 17–39. University of Chicago Press.

Friedman, Milton. 1967. "Must We Choose Between Inflation and Unemployment?" *Stanford Graduate School of Business Bulletin* 35(Spring): 10–13, 40, 42.

Friedman, Milton. 1968. "The Role of Monetary Policy." *American Economic Review* 58(1): 1–17.

Friedman, Milton. 1972. "Monetary Policy."

Proceedings of the American Philosophical Society 116(3): 183–96.

Friedman, Milton. 1975. *Unemployment versus Inflation? An Evaluation of the Phillips Curve.* London: The Institute of Economic Affairs.

Friedman, Milton. 1977. "Nobel Lecture: Inflation and Unemployment." *Journal of Political Economy* 85(3): 451–72.

Haberler, Gottfried. 1951. "Wage Policy, Employment, and Economic Stability." In *The Impact of the Union,* edited by David McCord Wright, 34–62. New York: Harcourt, Brace and Company.

Haberler, Gottfried. 1969. "Wage-Push Inflation Once More." In *Roads to Freedom: Essays in Honour of Friedrich A. von Hayek,* edited by Erich Streissler, Gottfried Haberler, Friedrich A. Lutz, and Fritz Machlup, 65–73. London: Routledge & Kegan Paul.

Haberler, Gottfried. 1972. "Incomes Policy and Inflation: Some Further Reflections." *American Economic Review* 62(2): 234–41.

Humphrey, Thomas M. 1977. "On Cost-Push Theories of Inflation in the Pre-War Monetary Literature." *Economic Review,* Federal Reserve Bank of Richmond, (May/June), 3–9.

Jacoby, Neil H. 1957. "Thinking Ahead." *Harvard Business Review* 35(3): 15–16, 20, 22–24, 26, 28, 30, 32, 160–62.

Joint Economic Committee. 1958. *Economic Policy Questionnaire: Tabulation of Replies Submitted to the Subcommittee on Economic Stabilization of the Joint Economic Committee, Congress of the United States, Eighty-Fifth Congress, Second Session.* December. Washington, DC: United States Government Printing Office.

Joint Economic Committee. 1959. "Statement of Milton Friedman, University of Chicago and National Bureau of Economic Research." In *Employment, Growth, and Price Levels: Hearings Before the Joint Economic Committee, Congress of the United States, Eighty-Sixth Congress, First Session, Pursuant to S. Con. Res. 13,* Part 4–The Influence on Prices of Changes in the Effective Supply of Money, 605–48. Washington, DC: United States Government Printing Office.

Lerner, Abba P. 1967. "Employment Theory and Employment Policy." *American Economic Review* 57(2): 1–18.

Lipsey, Richard G. 1960. "The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1862–1957: A Further Analysis." *Economica* 27(105): 1– 31.

Lipsey, Richard G. 1961. "Is Inflation Explosive?" *The Banker,* October, pp. 2–12.

Lipsey, Richard G. 1975. *An Introduction to Positive Economics,* 4th edition. London: Weidenfeld and Nicolson. (First edition 1963).

Machlup, Fritz. 1960. "Another View of Cost-Push and Demand-Pull Inflation." *Review of Economics and Statistics* 42(2): 125–39.

Miles, David, Ugo Panizza, Ricardo Reis, and Ángel Ubide. 2017. *And Yet It Moves: Inflation and the Great Recession.* Geneva Reports on the World Economy 19. Geneva and London: International Center for Monetary and Banking Studies (ICMB) and Centre for Economic Policy Research (CEPR).

Morton, Walter A. 1950. "Trade Unionism, Full Employment and Inflation." *American Economic Review* 40(1): 13–39.

Nelson, Edward. 2007. "Milton Friedman and U.S. Monetary History: 1961–2006." *Review,* Federal Reserve Bank of St. Louis, 89(3): 153–82.

Nelson, Edward. 2008. "Friedman and Taylor on Monetary Policy Rules: A Comparison." *Review,* Federal Reserve Bank of St. Louis, 90(2): 95–116.

Nelson, Edward. 2009. "An Overhaul of Doctrine: The Underpinning of UK Inflation Targeting." *Economic Journal* 119(538): F333–F368.

Nelson, Edward. Forthcoming. *Milton Friedman and Economic Debate in the United States, 1932–1972,* Book A. (Page numbers are from a draft on the author's website: https://sites.google.com/site/edwardnelsonresearch/.)

Nelson, Edward, and Anna J. Schwartz. 2008. "The Impact of Milton Friedman on Modern Monetary Economics: Setting the Record Straight on Paul Krugman's 'Who Was Milton Friedman?'" *Journal of Monetary Economics* 55(4): 835–56.

Phelps, Edmund S. 1967. "Phillips Curves, Expectations of Inflation and Optimal Unemployment Over Time." *Economica* 34(135): 254–81.

Phelps, Edmund S. 1968. "Money-Wage Dynamics and Labor-Market Equilibrium." *Journal of Political Economy* 76(4): 678–711.

Phelps, Edmund S. 2017. "Nothing Natural About the Natural Rate of Unemployment." *Project Syndicate,* November 2. https://www.project-syndicate.org/commentary/low-unemployment-subdued-inflation-paradox-by-edmund-s–phelps-2017-11.

Phillips, A. W. 1958. "The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957." *Economica 25*(100): 283–99.

Reder, Melvin W. 1952. "The Theory of Union Wage Policy." *Review of Economics and Statistics* 34(1): 34–45.

Reuber, Grant L. 1962. *The Objectives of Monetary Policy.* Working Paper Prepared for the Royal Commission on Banking & Finance. Ottawa: Queen's Printer.

Samuelson, Paul A. 1958. *Economics: An Introductory Analysis,* 4th edition. New York: McGraw-Hill Book Company. First edition 1948.

Samuelson, Paul A. 1961. *Economics: An Introductory Analysis,* 5th edition. New York: McGraw-Hill Book Company.

Samuelson, Paul A. 1970. *Economics,* 8th edition. New York: McGraw-Hill Book Company.

Samuelson, Paul A., and Robert M. Solow. 1960. "Analytical Aspects of Anti-Inflation Policy." *American Economic Review* 50(2): 177–94.

Schwarzer, Johannes A. 2012. "A. W. Phillips and His Curve: Stabilisation Policies, Inflation Expectations and the 'Menu of Choice.'" *European Journal of the History of Economic Thought* 19(6): 976–1003.

Schwarzer, Johannes A. 2014. "Growth as an Objective of Economic Policy in the Early 1960s: The Role of Aggregate Demand." *Cahiers d'économie politique/Papers in Political Economy* no. 67, pp. 175–206.

Schwarzer, Johannes A. 2016. *Price Stability versus Full Employment: The Phillips Curve Dilemma Reconsidered*. Stuttgart-Hohenheim: PhD Thesis, University of Hohenheim.

Slichter, Sumner H. 1952. "How Bad Is Inflation?" *Harper's Magazine,* August, 205(1227): 53–57.

Slichter, Sumner H. 1954. "Do the Wage-Fixing Arrangements in the American Labor Market Have an Inflationary Bias?" *American Economic Review* 44(2): 322–46.

Smithies, Arthur. 1957. "The Control of Inflation." *Review of Economics and Statistics* 39(3): 272–83.

Solow, Robert M. 1966. "The Case Against the Case Against the Guideposts." In *Guidelines, Informal Controls, and the Market Place: Policy Choices in a Full Employment Economy,* edited by George P. Shultz and Robert Z. Aliber, 41–54. University of Chicago Press.

Solow, Robert M. 1969. *Price Expectations and the Behavior of the Price Level: Lectures Given in the University of Manchester.* Manchester University Press.

Staff of the Cabinet Committee on Price Stability. 1969. *Studies by the Staff of the Cabinet Committee on Price Stability*. Washington, DC: United States Government Printing Office.

Taylor, John B. 2001. "An Interview with Milton Friedman." *Macroeconomic Dynamics* 5(1): 101–31.

Tobin, James. 1967. "Unemployment and Inflation: The Cruel Dilemma." In *Prices: Issues in Theory, Practice, and Public Policy*, edited by Almarin Phillips and Oliver E. Williamson, 101–107. Philadelphia: University of Pennsylvania Press.

Tobin, James. 1972. "Inflation and Unemployment." *American Economic Review* 62(1): 1–18.

Yellen, Janet L. 2016. "Macroeconomic Research after the Crisis." Speech at "The Elusive 'Great' Recovery: Causes and Implications for Future Business Cycle Dynamics" 60th Annual Economic Conference sponsored by the Federal Reserve Bank of Boston, Boston, Massachusetts, October 14, 2016.

Yellen, Janet L. 2017a. "Inflation, Uncertainty, and Monetary Policy." Speech at the "Prospects for Growth: Reassessing the Fundamentals" 59th Annual Meeting of the National Association for Business Economics, Cleveland, Ohio, September 26, 2017.

Yellen, Janet L. 2017b. "The U.S. Economy and Monetary Policy." Speech at the Group of 30 International Banking Seminar, Washington, DC, October 15, 2017.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 211–218

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., St. Paul, MN 55105.

Potpourri

David Miles, Ugo Panizza, Ricardo Reis, and Ángel Ubide have co-authored "And Yet It Moves: Inflation and the Great Recession." The authors pose a hypothetical question: If you were thinking about the path of inflation back in 2007 or so, and someone accurately described to you what was about to happen in the economy, what inflation rate would you have predicted? They write: "[G]iven how volatile and often high inflation has been in the past, given that there was a deep recession and brief deflation episode in 2008–10, given that nominal interest rates were virtually constant (and the real interest rate was not), given that the monetary base increased five-fold, and given that central banks undertook unprecedented policies in a context of fiscal volatility, what would have been your guess

■ *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.32.1.211

doi=10.1257/jep.32.1.211

about the stability and volatility of inflation from 2010 onwards? Simple versions of dominant economic theories, or superficial readings of economic history, would have all pointed to the conclusion that inflation should have at least been volatile, and possibly drifted up or down. Yet inflation was low and relatively stable. We did not observe deflation even in the presence of massive macroeconomic shocks and a sudden rise in unemployment, nor the much-feared inflation spiral that many expected after unprecedented easing in monetary policy. It is remarkable that the volatility of inflation remained so low, in spite of new policies and many shocks. … We will suggest that the stability of inflation poses puzzles for our existing theories, suggesting that inflation control is far from a solved problem. ... The young, or those with short memories, could be forgiven for looking condescendingly at their older friends who speak of inflation as a major economic problem. But, like Galileo Galilei told his contemporaries who thought the Earth was immovable, "Eppur si muove" ("and yet it moves"). …Will the great anchoring soon be followed by a great bout of inflation, or by a descent into deflation, just as the Great Moderation was followed by the Great Recession?" Geneva Reports on the World Economy 19, International Center for Monetary and Banking Studies and the Centre for Economic Policy Research, October 2017, http://voxeu.org/content/and-yet-it-moves-inflation-and-great-recession (with free registration).

The *World Development Report 2018* focuses on the theme "LEARNING to Realize Education's Promise." "The number of years of schooling completed by the average adult in the developing world more than tripled from 1950 to 2010, from 2.0 to 7.2 years. By 2010 the average worker in Bangladesh had completed more years of schooling than the typical worker in France in 1975. … By 2008 the average low-income country was enrolling students in primary school at nearly the same rate as the average high-income country. But schooling is not the same as learning. Children learn very little in many education systems around the world: even after several years in school, millions of students lack basic literacy and numeracy skills. In recent assessments in Ghana and Malawi, more than four-fifths of students at the end of grade 2 were unable to read a single familiar word such as *the* or *cat* … When grade 3 students in Nicaragua were tested in 2011, only half could correctly solve 5 + 6." "When improving learning becomes a priority, great progress is possible. In the early 1950s, the Republic of Korea was a war-torn society held back by very low literacy levels. By 1995 it had achieved universal enrollment in high-quality education through secondary school. Today, its young people perform at the highest levels on international learning assessments. Vietnam surprised the world when the 2012 results of the Programme for International Student Assessment (PISA) showed that its 15-year-olds were performing at the same level as those in Germany—even though Vietnam was a lower-middle-income country. Between 2009 and 2015, Peru achieved some of the fastest growth in overall learning outcomes—an improvement attributable to concerted policy action. In Liberia, Papua New Guinea, and Tonga, early grade reading improved substantially within a very short time thanks to focused efforts based on evidence." World Bank, 2018, http://www.worldbank.org/en/publication/wdr2018.

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel for 2017 was awarded to Richard Thaler "for his contributions to behavioural economics." The Nobel foundation has published the short and readable "popular information" essay "Easy Money or a Golden Pension? Integrating Economics and Psychology." It also published a longer "advanced information" essay, "Richard H. Thaler: Integrating Economics with Psychology." From this latter essay: "Richard Thaler played a crucial role in the development of behavioral economics over the past four decades. He provided both conceptual and empirical foundations for the field. By incorporating new insights from human psychology into economic analysis, he has provided economists with a richer set of analytical and experimental tools for understanding and predicting human behavior. This work has had a significant cumulative impact on the economics profession; it inspired a large number of researchers to develop formal theories and empirical tests, which helped turn a somewhat controversial, fringe field into a mainstream area of contemporary economic research. … In his well-known "Anomalies" series in the *Journal of Economic Perspectives*, as well as in many other articles, comments, and books, he continued to document and analyze how economic decisions are influenced by three aspects of human psychology: cognitive limitations (or bounded rationality), self-control problems, and social preferences. We organize this overview of Thaler's contributions around these three topics." The Nobel Committee essays are at https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2017/. A link to all of Thaler's "Anomalies" columns in this journal is at https://www.aeaweb.org/journals/jep/search-results?within%5Btitle%5D=on&within%5Babstract%5D=on&within%5Bauthor%5D=on&journal=3&q=Anomalies%3A.

Symposia

Cityscapes has published a 15-contribution symposium on "The Family Options Study." From the introduction by Anne Fletcher and Michelle Wood, "Next Steps for the Family Options Study": "HUD launched the Family Options Study in 2008 to learn about which housing and services interventions work best for families with children experiencing homelessness. Recruitment took place in emergency shelters across the 12 participating study sites. … In total, the study team enrolled 2,282 families, including nearly 5,400 children, into the study between September 2010 and January 2012. The study team followed the families for 3 years …" "The results of the Family Options Study offer striking evidence of the power of offering a long-term rent subsidy to a homeless family in shelter, substantially increasing housing stability and yielding benefits across a number of important domains, including reductions in residential moves, child separations, adult psychological distress, experiences of intimate partner violence, food insecurity, and school mobility among children, although those benefits were accompanied by reductions in work effort. These findings provide support for the notion that family homelessness is largely an economic issue, and that, by solving the economic issue, families experience

additional benefits that extend beyond housing stability. Equally notable is the fact that these significant benefits that accrued to the families offered a long-term rent subsidy were achieved at a comparable cost to other interventions tested, which offered few positive outcomes for families in any domain." US Department of Housing and Urban Development, 2017, vol. 19, no. 3, https://www.huduser.gov/portal/periodicals/cityscpe/vol19num3/index.html.

Diane Whitmore Schanzenbach and Ryan Nunn have edited an e-book of 11 readable essays: *The 51%: Driving Growth through Women's Economic Participation*. From the first essay by Sandra E. Black, Diane Whitmore Schanzenbach, and Audrey Breitwieser, titled "The Recent Decline in Women's Labor Force Participation": "[B]etween 1962 and 2000, women's labor force participation—defined as the percentage of women ages 16 and older either working or actively looking for work—increased dramatically, from 37 percent to 61 percent. … Estimates suggest that the economy is $2.0 trillion, or 13.5 percent, larger than it would have been had women's participation and hours worked remained at their 1970 levels. … However, beginning in 2000, the positive trends slowed and even reversed: women's participation fell from 60.7 percent in 2000 to 57.2 percent in 2016. … Why has the progress stopped, and even reversed … Importantly, the United States appears to be an outlier in terms of women's labor force participation; France, Canada, the United Kingdom, and Japan all continued to see positive growth in prime-age women's labor force participation post-2000, with levels rising substantially above those in the United States. This divergence suggests a significant role for labor-market institutions." Hamilton Project at the Brookings Institution, October 2017, https://www.brookings.edu/wp-content/uploads/2017/10/es_101917_the51percent_full_book.pdf.

The *Cato Journal* has published an 11-paper symposium on "The Economics of Immigration." For example, Giovanni Peri writes about "The Impact of Immigration on Wages of Unskilled Workers": "Immigrants did not contribute to the national decline in wages at the national level for native-born workers without a college education. This article reviews how the timing of their immigration and skill sets of immigrants between 1970 and 2014 could not have been responsible for wage declines. This article then reviews other evidence at the local level that implies immigration is not associated with wage declines for noncollege workers, even if they are high school dropouts. Higher immigration is associated with higher average wages. Causality is difficult to tease out but numerous factors could explain the positive association between the quantity of immigrants and native wages." Cato Institute, Fall 2017, https://www.cato.org/cato-journal/fall-2017.

Lectures

Larry Summers delivered a speech on "Rethinking Global Development Policy for the 21st Century" at the annual Global Development Changemaker Dinner of the Center for Global Development. "[B]etween the time of Pericles and London

in 1800, standards of living rose about 75 percent in 2,300 years. They called it the Industrial Revolution because for the first time in human history, standards of living were visibly and meaningfully different at the end of a human lifespan than they had been at the beginning of a human lifespan, perhaps 50 percent higher during the Industrial Revolution. Fifty percent is the growth that has been achieved in a variety of six-year periods in China over the last generation and in many other countries, as well. And so if you look at material standards of living, we have seen more progress for more people and more catching up than ever before. That is not simply about things that are material and things that are reflected in GDP. ... [I]f current trends continue, with significant effort from the global community, it is reasonable to hope that in 2035 the global child mortality rate will be lower than the US child mortality rate was when my children were born in 1990. That is a staggering human achievement. It is already the case that in large parts of China, life expectancy is greater than it is in large parts of the United States." November 8, 2017. Text of the 45-minute lecture is at https://www.cgdev.org/sites/default/files/Rethinking-Global-Development-Policy-for-21st-Century.pdf, while video is at https://www.cgdev.org/event/rethinking-global-development-policy-21st-century.

Peter H. Lindert delivered the OECD Angus Maddison Development Lecture on "The Rise and Future of Progressive Redistribution" on October 3, 2017. From the abstract of his background paper of the same title: "There appears to have been a global shift toward progressive redistribution over the last hundred years in all prosperous countries. The retreats toward regressive redistribution have been rare and have been reversed. As a corollary, the rise in income inequality since the 1970s owes nothing to any retreat from progressive government spending. Adding the effects of rising subsidy for public education on the later inequality of adult earning power strongly suggests that a fuller, longer-run measure of fiscal incidence would reveal a history of still greater shift toward progressivity, most notably in Japan, Korea, and Taiwan. The key determinant of progressivity in the decades ahead is population aging, not inequality itself or immigration backlash." Tulane University, Commitment to Equity Institute, October 2017, Working Paper 73, http://www.commitmentoequity.org/wp-content/uploads/2017/11/CEQ-WP73_Lindert_Rise-FutureProgressiveRedistribution_Oct17_2017.pdf.

Alan B. Krueger delivered the annual Daniel Patrick Moynihan Lecture on Social Science and Public Policy, on the topic of "Independent Workers: What Role for Public Policy," for the American Academy of Political and Social Science: "One policy proposal that has gained some traction is to have a carve out for intermediaries that permits them to provide benefits without risk that their contractors will be deemed employees." "For the self-employed, however, health insurance expenses are excluded from income taxes but not from payroll taxes. With payroll taxes of around 15 percent, this creates a significant additional tax on the self-employed. That could easily be rectified through tax policy. As mentioned, the self-employed receive relatively little job training. The IRS is tough on the deductibility of training expenses for the self-employed. Particularly when it comes to safety-related training, it would make sense for the IRS to be more permissive in allow training deductions

as a business expense. Congress could also enact tax credits to encourage job training, particularly for safety training, for self-employed workers." "*Extend coverage under Title VII of the Civil Rights Act of 1964 to independent contractors.* The self-employed currently have few options if they face discrimination." "Here's a really ambitious, big idea: … 'Shared Security Accounts,' in which all workers would be covered by a universal system that provides health insurance, retirement benefits, paid leave, and so on. Employers and online platforms like Uber would contribute 25% of their workers' compensation into a fund to pay for those benefits. Workers could choose which benefits they want. ... Washington State and New Jersey have considered legislation along these lines for self-employed workers." Video of the hour-long lecture delivered on May 18, 2017, is http://www.aapss.org/news/alan-krueger-delivers-2017-moynihan-lecture/. A revised and written-out version of the lecture is available as Princeton University Industrial Relations Section Working Paper 615, September 2017, http://dataspace.princeton.edu/jspui/bitstream/88435/dsp01tt44pq514/3/615.pdf.

Mervyn King delivered the 2017 Martin Feldstein Lecture at the National Bureau of Economic Research on the subject of "Uncertainty and Large Swings in Activity." "Imagine that you had a problem in your kitchen, and summoned a plumber. You would hope that he might arrive with a large box of tools, examine carefully the nature of the problem, and select the appropriate tool to deal with it. Now imagine that when the plumber arrived, he said that he was a professional economist but did plumbing in his spare time. He arrived with just a single tool. And he looked around the kitchen for a problem to which he could apply that one tool. You might think he should stick to economics. But when dealing with economic problems, you should also hope that he had a box of tools from which it was possible to choose the relevant one. And there are times when there is no good model to explain what we see. The proposition that 'it takes a model to beat a model' is rather peculiar. Why does it not take a fact to beat a model? And although models can be helpful, why do we always have to have one? After the financial crisis, a degree of doubt and skepticism about many models would be appropriate." A written version of the presentation is available in the *NBER Reporter*, no. 3, September 2017, pp. 1–10), at http://www.nber.org/reporter/2017number3/2017number3.pdf, and video of the lecture delivered on July 19, 2017, is at http://www.nber.org/feldstein_lecture_2017/feldstein_lecture_2017.html.

Interviews

Douglas Clement has an "Interview with Lawrence Katz," with the subheading: "Harvard economist on the gender pay gap, fissuring workplaces and the importance of moving to a good neighborhood early in a child's life." "If you look at the past 30 years, … we estimate that, as recently as 2013, about two-thirds of that [increase in inequality] is due to the growth of the educational wage premium. … So, if you'd kept the college premium at the 1980 level, you would've seen only a

third as much of the growth of U.S. earnings inequality. ... What the government has done—in the '50s and '60s, even into the '70s—is invested heavily in high-quality colleges. Think of University of California campuses or Florida State. But since then, there's been very little investment in expanding quality higher education. There's increased crowding at community colleges and state universities, and states have greatly cut back on appropriations for higher education, particularly in the Great Recession. The federal government has continued to have an important role, but it's done it with flexible support through Pell grants targeted to low-income students. The problem is that we've had a surge of really low-quality colleges, and the worst of that is the for-profit sector … It's been a bit of a disaster. Even though these for-profit institutions have tried to be up to date, very flexible, with high-quality online instruction, we have repeatedly found very little economic return to degree programs at for-profit institutions; instead, it's become a massive debt trap." *The Region*, Federal Reserve Bank of Minneapolis, September 25, 2017, at https://www.minneapolisfed.org/publications/the-region/interview-with-lawrence-katz.

In an "Interview" with Jesse Shapiro, Renee Haltom elicits insights on topics related to media and political bias. On ideological segregation: "Take the fraction of the audience on a given news site that is conservative and call that the conservativeness of the site. Then take the website visited by the average conservative on the average day—that website is about as conservative as usatoday.com. Now do that same thing for the average liberal, that's about as liberal as cnn.com. If you were to read those two outlets, you wouldn't find that they're radically different. In fact, we find that isolation is very rare in the data. … The people who are consuming niche media are probably pretty politically engaged people, and therefore they want to read a lot of things. So in the end, the picture is a lot more muted than what people have feared." On social media and polarization: "Our favorite and most important comparison is with respect to age. People who are 75 years and over rarely use social media and don't report getting a lot of political information online. People who are 18 to 25 frequently use social media and report getting a lot of political information online. So if you thought that social media was contributing to the rise in polarization, what you would expect to see in the data is that polarization is rising especially fast for younger Americans — and if anything, the story is the opposite. … I think the effect of the Internet on polarization remains an open question." *Econ Focus*, Federal Reserve Bank of Richmond, 2nd Quarter 2017, pp. 24-29, https://www.richmondfed.org/publications/research/econ_focus/2017/q2/interview.

Discussion Starters

"*The Lancet* Commission on Pollution and Health," convened by the British medical journal, included about four-dozen members. "Diseases caused by pollution were responsible for an estimated 9 million premature deaths in 2015—16% of all deaths worldwide—three times more deaths than from AIDS, tuberculosis, and malaria combined and 15 times more than from all wars and other forms of

violence. In the most severely affected countries, pollution-related disease is responsible for more than one death in four. Pollution disproportionately kills the poor and the vulnerable. Nearly 92% of pollution-related deaths occur in low-income and middle-income countries and, in countries at every income level, disease caused by pollution is most prevalent among minorities and the marginalised." October 19, 2017, available with free registration at http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(17)32345-0.pdf.

The International Labour Organization has published "Global Estimates of Child Labour: Results and Trends 2012–2016." "The challenge of ending child labour remains formidable. A total of 152 million children—64 million girls and 88 million boys—are in child labour globally, accounting for almost one in ten of all children worldwide. Nearly half of all those in child labour—73 million children in absolute terms—are in hazardous work that directly endangers their health, safety, and moral development. Children in employment, a broader measure comprising both child labour and permitted forms of employment involving children of legal working age, number 218 million." September 2017, http://www.ilo.org/wcmsp5/groups/public/—dgreports/—dcomm/documents/publication/wcms_575499.pdf.

Jesse Bricker, Lisa J. Dettling, Alice Henriques, Joanne W. Hsu, Lindsay Jacobs, Kevin B. Moore, Sarah Pack, John Sabelhaus, Jeffrey Thompson, and Richard A. Windle discuss "Changes in U.S. Family Finances from 2013 to 2016: Evidence from the Survey of Consumer Finances." On the distribution of wealth: "The wealth share of the top 1 percent climbed from 36.3 percent in 2013 to 38.6 percent in 2016, slightly surpassing the wealth share of the next highest 9 percent of families combined … After rising over the second half of the 1990s and most of the 2000s, the wealth share of the next highest 9 percent of families has been falling since 2010, reaching 38.5 percent in 2016. Similar to the situation with income, the wealth share of the bottom 90 percent of families has been falling over most of the past 25 years, dropping from 33.2 percent in 1989 to 22.8 percent in 2016." *Federal Reserve Bulletin* Summer 2017, pp. 1–42, at https://www.federalreserve.gov/publications/files/scf17.pdf.

J. Bradford DeLong has written "When Globalization Is Public Enemy Number One" in the most recent issue of the *Milken Institute Review*. "To repeat, because it bears repeating: globalization in general and the rise of the Chinese export economy have cost some blue-collar jobs for Americans. But globalization has had only a minor impact on the long decline in the portion of the economy that makes use of high-paying blue-collar labor traditionally associated with men. ... Pascal Lamy, the former head of the World Trade Organization, likes to quote China's sixth Buddhist patriarch: 'When the wise man points at the moon, the fool looks at the finger.' Market capitalism, he says, is the moon. Globalization is the finger." *Milken Institute Review* Fourth Quarter 2017, pp. 22–31, http://www.milkenreview.org/articles/when-globalization-is-public-enemy-number-one.

Journal of Economic Perspectives—Volume 32, Number 1—Winter 2018—Pages 219–220

Using JEP Articles as Course Readings? Tell Us About It!

Probably the most common metric for judging the impact of an academic journal on the economic profession is the number of times that articles in that journal are cited. The *Journal of Economic Perspectives* does just fine by this measure. For example, according to the *Journal Citation Reports,* published by Thomson Reuters, the JEP ranked third in 2016 in "journal impact factor" among the 347 economic journals in the database. Thus, the articles published in the JEP seem to be widely cited and, by implication, widely read among research economists.

But while research economists are the core constituency for JEP, this journal is a little different from standard refereed journals in that it aims to serve several additional audiences. For example, we hope that some of the articles in the journal reach out to the policy community and inform the public debate on important economic policies. We also hope that JEP articles are useful for teaching and for students. In particular, we hope that our papers end up on reading lists and syllabuses, especially for undergraduate courses, and provide useful background material for lectures or seminars, or starting points to recommend to inquisitive students.

In this spirit, if you are have JEP articles on your syllabus, we earnestly request your assistance. To facilitate and foster the use of JEP articles in the classroom, we would like to collect and make available concrete examples of successful use of JEP articles on reading lists or in classroom settings.

This invitation is meant broadly. If you are just using one or a few JEP articles in the classroom, and they are working well for you, let us know. If you sometimes assign JEP articles to groups of students and then have the students explain the articles to the rest of the class, tell us about it. If you are running a JEP-centric class with a substantial proportion of JEP articles on the reading list, we definitely want to hear from you. Our main focus is on undergraduate courses, but if you have recommendations at the graduate level, we are glad to hear about those, as well.

If time and energy permit, we would also appreciate your typing out a few lines to let us know how long the articles have been on your reading list, and to give us a sense of what articles are working best for you and your students. Please feel encouraged to attach a copy of your syllabus, too.

What we do with the answers at this end will depend on the magnitude and details of the response we receive. Ideally, assuming a reasonable number of responses to this note, we would compile a relatively short article in JEP that would describe the response, and list some of the most widely-used articles for different

courses. Along with that article, we could post on the JEP webpage a more detailed description, course by course, of what JEP articles are being used successfully. We could also offer some testimonial evidence from those who have used them.

If you would like to share your JEP-related class material, please send an email to Timothy Taylor, Managing Editor of JEP, at taylort@macalester.edu. If you know of colleagues who use JEP material in their classes, please help us in spreading the word.

Thank you for your help.

Enrico Moretti, Editor
Timothy Taylor, Managing Editor

ADVANCING KNOWLEDGE THROUGH DATA AND RESEARCH

HUD User is our nation's premier source of housing research and data. Visit **www.huduser.gov** to access free research publications and data sets. Browse our eBookstore for our most popular housing research in eBook-friendly formats, and download our free apps to quickly obtain data and information on your mobile device. Subscribe to receive email updates through our eLists and check out our recently redesigned Datasets webpage at **www.huduser.gov**, where you can easily find what you need.

U.S. DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT

PD&R U.S. Department of Housing and Urban Development Office of Policy Development and Research

HUD USER

ECONOMICS RESEARCH STARTS HERE

Easily Access An Essential Economics Library:

- ***>> 1.5 million bibliographic records spanning 130 years, with nearly 70,000 additions per year***
- ***>> Optional full-text of over 500 economics journals including all journals published by the American Economic Association***
- ***>> Indexes of journal articles, working papers, PhD dissertations, book reviews, conference proceedings, and collective volume articles***
- ***>> International coverage includes journals published in 74 countries***

AMERICAN ECONOMIC ASSOCIATION

EconLit™
www.econlit.org

AEA's

JOE NETWORK

The Preferred Hiring Tool for the Economics Job Market

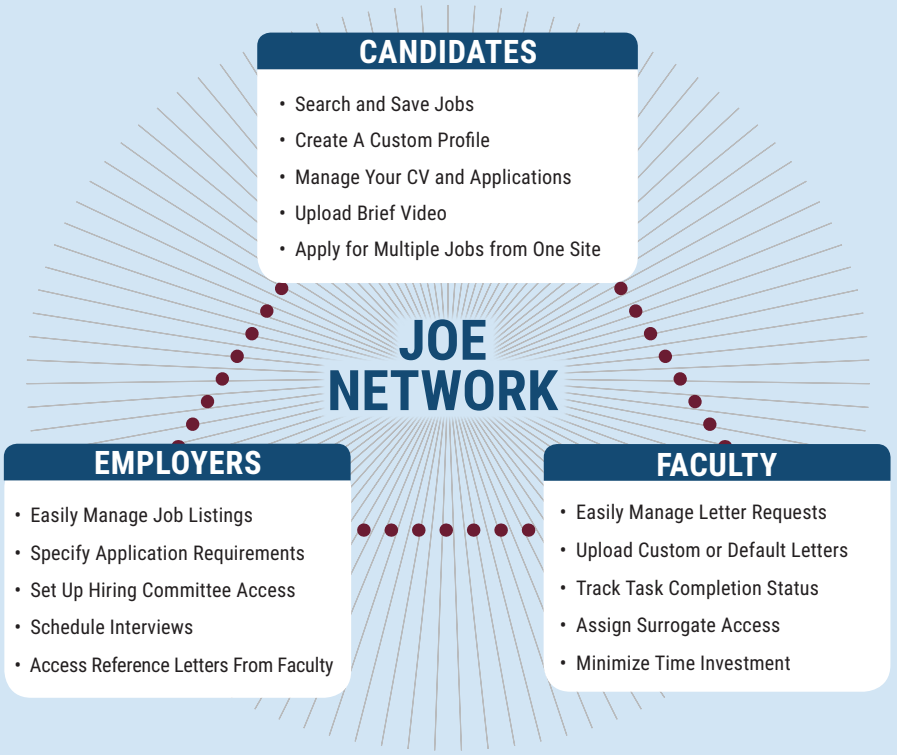
Employers	Candidates	Faculty
1,700+ Positions filled	**5,000+** From around the world	**142,000+** Reference letter requests

The JOE Network is designed for economists…by economists.

No other placement program offers a more comprehensive way to match high-caliber candidates with sought after economics positions—including secure management of the faculty reference letter writing process.

JOE Network automates all hiring tasks. Users share materials, communicate confidentially, and easily manage their files and data right from the AEA website.

Automates the Economics Job Market Process for All Participants



Great Jobs. Great Candidates. Great Careers.

JOE Network is the Preferred Hiring Tool for the Economics Job Market

AMERICAN ECONOMIC ASSOCIATION

Learn more about the JOE Network at the AEA's website!

www.aeaweb.org/JOE

DON'T MISS…

CTREE 2018

The Eighth Annual Conference on Teaching & Research in Economic Education

Plenary Speakers include:

Sandra Black
University of Texas

Catherine Eckel
Texas A&M University

Dan Hamermesh
University of Texas

May 30 – June 1, 2018
San Antonio, Texas
San Antonio Marriott Rivercenter

CTREE is hosted by the AEA Committee on Economic Education in conjunction with the *Journal of Economic Education*.

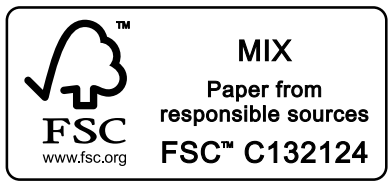
Early-bird registration opens February 15, 2018

For more info go to
www.aeaweb.org/ctree/2018

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.

Copyright © 2018 by the American Economic Association. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation, including the name of the author. Copyrights for components of this work owned by others than AEA must be honored. Abstracting with credit is permitted. The author has the right to republish, post on servers, redistribute to lists, and use any component of this work in other works. For others to do so requires prior specific permission and/or a fee. Permissions may be requested from the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203; e-mail: aeainfo@vanderbilt.edu.



Founded in 1885

EXECUTIVE COMMITTEE

Elected Officers and Members

President
OLIVIER BLANCHARD, Peterson Institute for International Economics
President-elect
BEN S. BERNANKE, The Brookings Institution
Vice Presidents
SUSAN C. ATHEY, Stanford University
PINELOPI KOUJIANOU GOLDBERG, Yale University
Members
JOHN Y. CAMPBELL, Harvard University
HILARY HOYNES, University of California at Berkeley
NICHOLAS BLOOM, Stanford University
ERICA FIELD, Duke University
ADRIANA LLERAS-MUNEY, University of California at Los Angeles
BETSEY STEVENSON, University of Michigan
Ex Officio Members
ROBERT J. SHILLER, Yale University
ALVIN E. ROTH, Stanford University

Appointed Members

Editor, The American Economic Review
ESTHER DUFLO, Massachusetts Institute of Technology
Editor, The American Economic Review: Insights
AMY FINKELSTEIN, Massachusetts Institute of Technology
Editor, The Journal of Economic Literature
STEVEN N. DURLAUF, University of Wisconsin
Editor, The Journal of Economic Perspectives
ENRICO MORETTI, University of California at Berkeley
Editor, American Economic Journal: Applied Economics
ALEXANDRE MAS, Princeton University
Editor, American Economic Journal: Economic Policy
MATTHEW D. SHAPIRO, University of Michigan
Editor, American Economic Journal: Macroeconomics
SIMON GILCHRIST, New York University
Editor, American Economic Journal: Microeconomics
JOHANNES HÖRNER, Yale University
Secretary-Treasurer
PETER L. ROUSSEAU, Vanderbilt University

OTHER OFFICERS

Editor, Resources for Economists
WILLLIAM GOFFE, Pennsylvania State University
Director of AEA Publication Services
JANE EMILY VOROS, Pittsburgh
Managing Director of EconLit Product Design and Content
STEVEN L. HUSTED, University of Pittsburgh
Counsel
TERRY CALVANI, Freshfields Bruckhaus Deringer LLP
Washington, DC

ADMINISTRATORS

Director of Finance and Administration
BARBARA H. FISER
Convention Manager
GWYN LOFTIS

The Journal of

Economic Perspectives

Winter 2018, Volume 32, Number 1

Symposia

Housing

Friedman's Natural Rate Hypothesis after 50 Years

Articles

Features

AMERICAN ECONOMIC ASSOCIATION